Trustworthy Artificial Intelligence in Healthcare: a proposed framework

Niamh St John Lynch1[0009-0009-4150-4970], Róisín Loughran2[0000-0002-0974-7106], Martin McHugh3[0000-0003-4275-3302], Fergal McCaffrey4[0000-0002-0839-8362]

1,2,3,4 Regulated Software Research Centre, DkIT, Dundalk, Co. Louth, Ireland

niamh.stjohnlynch@dkit.ie

**Abstract**: Trustworthy Artificial Intelligence (TwAI) is not sufficiently defined within healthcare to enable legal manufacturers and developers to identify and operationalise, with a high degree of assurance, the General Safety and Performance Requirements. This includes patients’ fundamental rights which are necessary to meet European regulatory requirements and stakeholder expectations. Medical Device Software developers are tasked with identifying appropriate performance measures for robustness of Artificial Intelligence-enabled Medical Devices based on the use case and risks identified by the development team. Risk identification is heavily reliant on the adequacy of their risk management process and the competency of personnel. As medical device software defects continue to be identified in the post-market phase, further research and guidance is necessary. Medical device manufacturers have reported a lack of understanding and harmonisation of the requirements that make up TwAI in healthcare, starting with ethical requirements. With standards development still in progress, research is necessary to ensure a common framework is applied, having appropriate development processes, methods, and measures for delivery of TwAI in healthcare. Therefore, we conducted a review of trust for two reasons: 1) to attempt to gain a collective understanding of TwAI in healthcare; 2) to propose a framework that considers the regulatory landscape for Artificial Intelligence-enabled Medical Devices that is sufficiently detailed to move beyond theoretical concepts. This research brings together the state of the art for artificial intelligence in healthcare for a harmonised approach and extends the literature by presenting a proposed framework to be operationalised by legal manufacturers and developers who place medical devices on the market for clinical use.

**Keywords**: Trustworthy, Medical Device Software Development, Artificial Intelligence, AI, Healthcare, Artificial Intelligence-enabled Medical Device, AI Development, Fundamental Rights, Ethics.

# 1 Introduction

This paper aims to review the current state of Trustworthy Artificial Intelligence (TwAI) within Medical Device Software (MDSW) and provide a framework for achieving TwAI in healthcare. To this end, a review of the current research and regulatory guidelines is performed with a focus on healthcare and AI-enabled Medical Device (AIeMD) technologies within the medical device sector.

It is well documented that there are considerable benefits promised from the increase in Artificial Intelligence (AI) in healthcare. These include improved clinical treatments and care, better diagnostic performance, improved elucidation of pathologies and specialised treatments closely matched to suit individual patients, amongst others [1], [2], [3]. These benefits go hand-in-glove with health concerns and challenges surrounding AIeMD, including but not limited to *data security* and *integrity*, *inadequate* or *biased data sets* used in training, *biased* or *inaccurate* *results*, *degradation*, and not least *autonomous decisions* and *autonomous change* of AI algorithms in use. All of these can lead to the potential for negative consequences for patient health and healthcare organisations. Recognition of the need for sufficient oversight and monitoring has been identified, though how this will be achieved has not yet been clearly defined [4]. Oversight being used as part of the intended use, indicating that AI is providing clinical support only is being utilised to lower the risk classification of AIeMDs for release to market. This is of particular concern in light of the risk of over-trust, which is likely to occur at some point, particularly in under-resourced and over-stretched healthcare systems. Lower risk classifications can result in less scrutiny placed on devices during regulatory review. Furthermore, based on the current version of the software development lifecycle (SDLC) standard IEC 62304, reduced software engineering rigor is applied in development and maintenance based on the software safety classification of A, B, and C; C being at the highest risk-level with potential for death or serious injury [5], [6].

The European Commission’s (ECs) report on AI Ethics Guidelines [7] for TwAI aims to provide a *high*-*level* framework. As stated within current research and noted within the guideline itself, there is plenty of room for interpretation and discretion, where the ethics guideline identifies a high-level agnostic framework only and accepts it is neither complete nor comprehensive. It is, nevertheless, a starting point for identifying methods for achieving TwAI [1]. There has been much research focusing on TwAI of late, though none have yielded a comprehensive TwAI for healthcare with sufficient detail to be useful [4], [8], [9]. Indeed, this is a gap known to the EC and representative standards committees such as CEN-CENELEC, who have in 2025 requested work commence to create six new standards in order to support the EU AI Act. One of the standards requested is the *AI Trustworthiness Framework* which forms part of the 2030 Strategy for CEN-CENELEC Standardisation [10]. The research performed here is to be utilised to support the creation of this standard by JTC 21, who have agreed it will be a Technical Report (TR) and available as guidance to industry. As a TR, this document cannot be used for compliance purposes though will be helpful in understanding the legislative requirements. The only other framework that deals with ethics in the manner identified here is that of the UN, which is not healthcare specific and again does not provide specific details for operationalisation in healthcare. The UN’s framework for ethical AI is concerned with the use of AI within the UN and is not focused on healthcare or AIeMD development activities. The most recent guidance released by the Medical Device Coordination Group (MDCG) and Joint AI Board [11] also indicate that fundamental rights must be assessed throughout the development lifecycle, but beyond that, indicate that standards are under development.

Section 2 looks at the literature on trust and what it means for trust of AI in healthcare. We present a brief overview of what is meant by TwAI in healthcare. We identify various key components that are suggested by scholars to form part of TwAI based on literature review. We then present a proposal for a more comprehensive and holistic perspective of TwAI in healthcare arising from this research based on literature review and semi-qualitative interview and expert reviews. The aim is to move closer towards a framework that can be operationalised within the existing SDLC process for AIeMDs. The framework proposed in this research is validated by expert review under International Standards Technical Committees and through expert reviews both on-line and during semi-qualitative interviews by leading AI developers.

Section 3 presents how the framework proposed in this research was developed and addresses the methodology employed; utilising both primary and secondary research methods. Section 4 provides a summary of three interconnected models proposed by this research, which together make up the full TwAI framework for healthcare before moving onto Sections 5 and 6, Discussion and Conclusion, respectively.

# 2 Ethical Basis of Trust

**2.1 What is Trust?**

For decades scholars have considered the concept of trust and what it means to be trustworthy [12], [13]. Many historical definitions of trust refer to social and human relationships that are not appropriate for this use case. It is accepted that ‘*trust’* is a complex and multi-dimensional phenomenon and difficult to operationalise, measure and interpret [14]. According to Simpson [14] *trust* is defined as “*confidence that [one] will find what is desired [from an-other] rather than what is feared*”, adopted from Deutsch 1973, p.148 [14]. We do not assume that *trust* in AI and TwAI are one and the same, but that they are intertwined and complex. Given the digital age we are in, we see concepts of trust being extended to a theory of computational trust [15] arguing that a theory of trust should mirror human expectations to the largest extent possible. Cognitive scientists in Italy, Castelfranchi & Falcone [16] present *trust* through the dynamic and complex lens of cognition, affective processes and social attributes, though admit it is difficult to pin-down. They present a socio-cognitive and computational model of trust theory for a deeper understanding of the *cognitive mediators* and *societal affects* following >10-years of discussions with cognitive scientists and AI developers [16]. We recognise the concept of trust is not new and is not intended to be covered at length here. What is necessary is for us, is to recognise that terms such as ‘*trust’*, ‘*trustworthy’* and ‘*trustworthiness’* are predominantly psychological constructs that are dynamic in nature and reflect a dynamic environment, with fluctuations dependent on the use case on which we place that trust [17], [18].

**2.3 Trust in AIeMD**

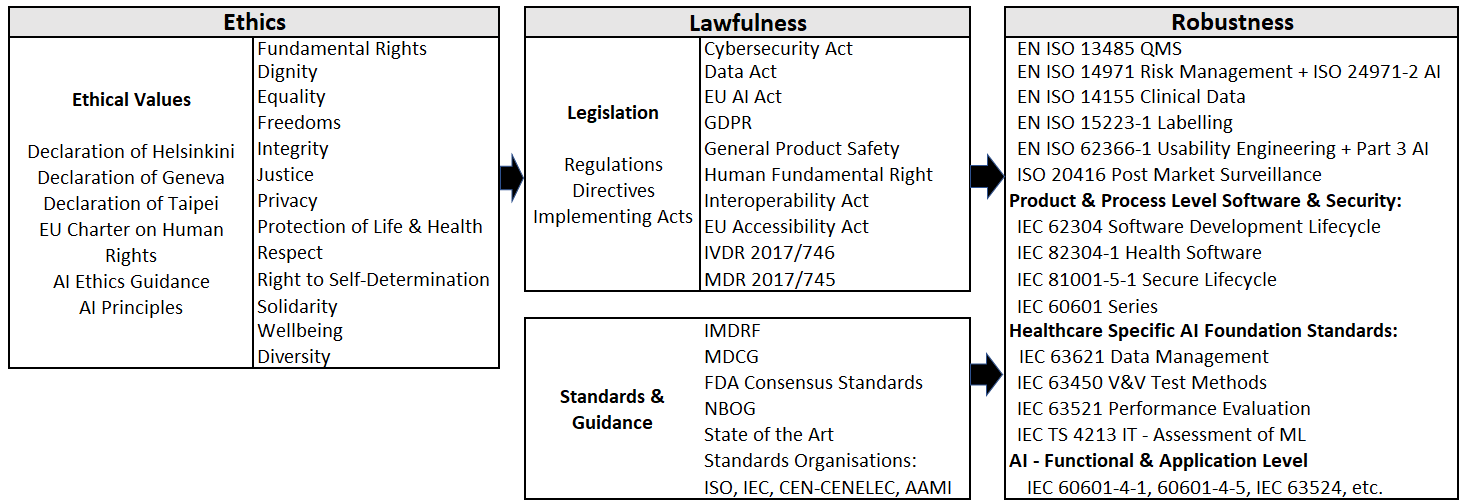
Within AIeMD research, Starke and colleagues identify two camps emerging in relation to trust in healthcare, 1) that *trust* in AI is conceptionally confused and that we *cannot* trust AI and 2) that we *should not* trust AI due to high stakes to health in the medical arena [19]. Starke et al. [19] go on to suggest that to place trust in AI is conceptionally plausible once we distinguish between reliability, competency and intention, noting the need for a definition of acceptable thresholds and regulatory scrutiny. Indeed, the most recent proposal for a trustworthiness framework put forward for ballot by CEN-CENELEC identifies key aspects such as *transparency*, *human oversight*, *accuracy* and *robustness* but these alone are piecemeal.

Some researchers call for the need to trust AI regardless of understanding its mechanisms of action, often referred to as black-box algorithms [20], [21]*.*  Moreover, Omrani, et al [20] suggest that to trust AI assumes an ethical basis of trust can be applied. It is important to understand the ethical and moral beliefs that go into generating agreements, such as the Declaration of Helsinki (DoH), Declaration of Geneva (DoG), Declaration of Taipei (DoT) and the EU Charter of Human Rights (CoHR) together with the WMA Declaration on Patient Safety (DoPS), in order to address Fundamental Rights called out in the EU AI Act. We cannot assume patient rights will be delivered where there is confusion as to how to achieve it. This research agrees with that of Zhang and Zhang who suggest, *ethical values* belong as a ‘*top-level design*’ [22], which is not the approach taken by the EU Ethics Guideline, that describe three pillars: Ethics, Lawfulness and Robustness.

**2.3.1 Ethical Values in Medical Devices.** We see evidence of ethical principles in medical research, though predominantly in clinical investigations that arise after development. For instance, the DoH on Ethical Principles for Medical Research Involving Human Subjects [23] is called out in the 1st Chapter of the Medical Device Regulations [24], Section 64. The DoH forms the basis of clinical investigations as defined in the MDR and IVDR and is further supported by relevant harmonised international standards, e.g., ISO-14155:2020 for clinical investigations of all medical devices [25]. ISO 20916:2019 [26] identifies ethical considerations during clinical studies, to ensure clinical investigations are designed and conducted in such a way that peoples’ rights, safety, dignity, and well-being of those participating in clinical performance studies are protected. Patient interests must prevail over all other interests with a guarantee that they are scientifically valid, reliable, and robust. Applied ethics in healthcare is not new and is well established, going back to the Hippocratic Oath and designed to keep clinicians accountable for care of patients [27]. However, ethics within the manufacture of medical devices through the development lifecycle is not yet widely considered or typically seen as a requirement within the SDLC, beyond usability risk assessments and clinical intended use capturing a representative population [28]. Indeed, the EU Ethics Guideline identifies AI ethics as a sub-field of applied ethics that is necessary for “*development, deployment and use of AI*” [7].

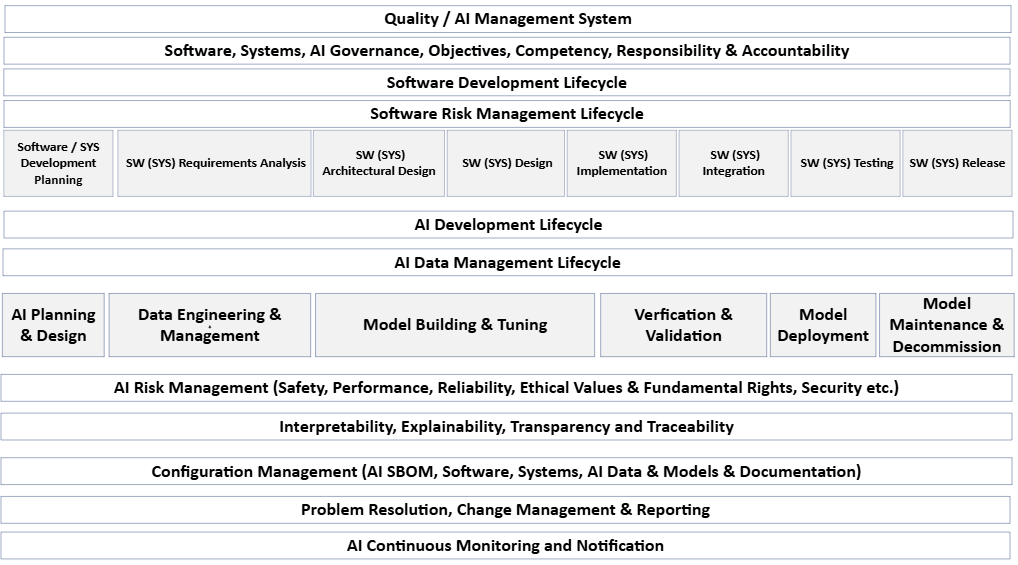
With the recent emergence of autonomous AI in healthcare, the need to better understand our agreed ethical values has become apparent. Also, the need to define ways of identifying these early within the SDLC need to be defined. Ethics have not been a consideration in traditional SDLCs, which has led this research to identify applicable ethical values necessary for consideration within an AI Development Lifecycle (AIDL). To this end, an Ethics, Lawfulness and Robustness (ELR) model for healthcare is proposed as part of this research, ref. **Figure 1**. The proposed ELR documents existing values and legislation and does so in a manner that has not been represented for this purpose before or in this comprehensive manner for application in AIeMD development.

The ELR model has been reviewed by TC62/MT49 Technical Committee (TC) as part of IEC 62304 (2nd Edition) draft standard. Experts within TC62/MT49 include medical device software and AI experts representing standards bodies globally (US, China, Japan, Australia, Europe, etc.). The review led to the removal of the ELR model from IEC 62304 ED2, where there was not full consensus for inclusion, within some suggesting ethics were not relevant for an SDLC. This demonstrates the disparity and changing landscape arising from AI and the work that needs to be done to move towards harmonising different regulatory perspectives. It was agreed by the TC and experts alike, that the ELR model was very helpful and should form part of a standard related to ethics specifically. This position was in contrast to many expert reviews performed during semi-quantitative interviews of AIeMD developers from leading organisations. Companies interviewed included Brain Lab, GE Healthcare, Siemens Healthineers and more; all leaders of AI development at this time. It is important that all AIeMD developers from small-medium sized organisations and large multinationals understand the specific ethical values to include in their risk assessments and identify appropriate requirements under an AI Development Lifecycle that this framework provides. The ELR is only part of the picture however, and the mechanisms to describe delivering safety, reliability and robust performance are also necessary through the AI-Development Lifecycle and Change Management and Monitoring Systems.



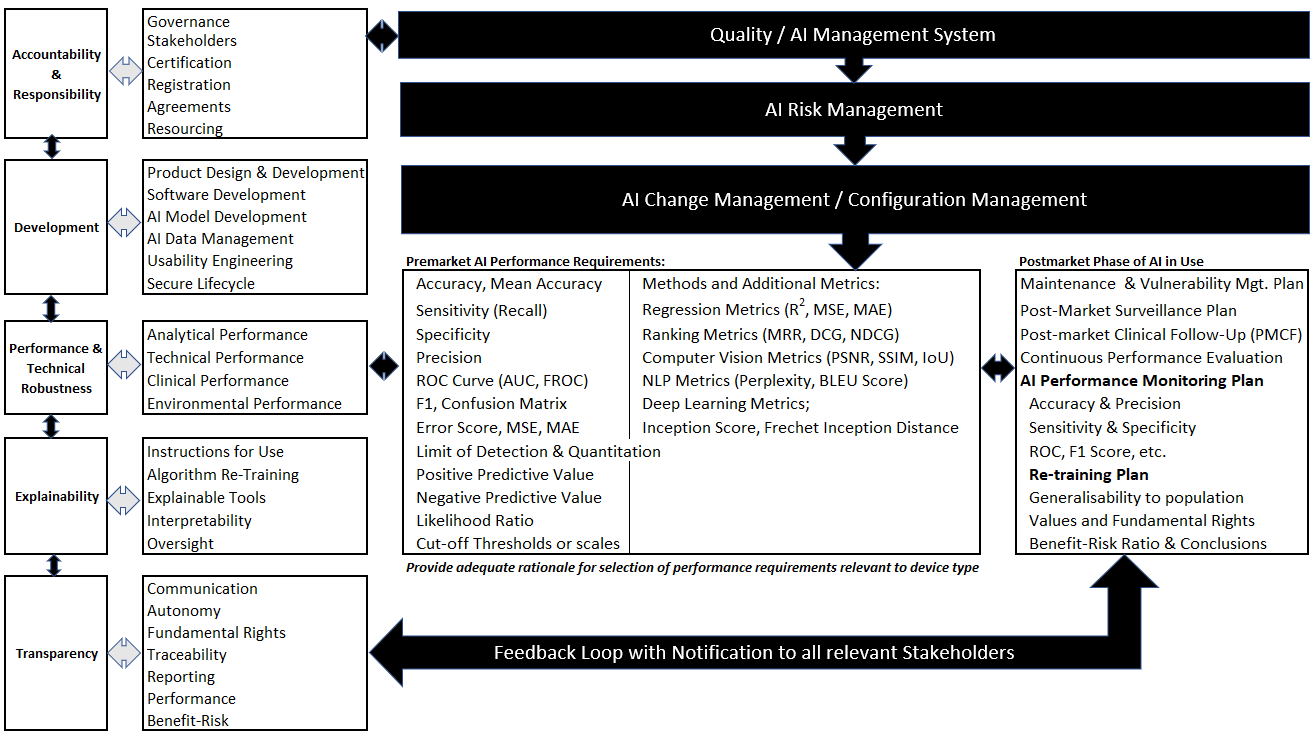
**Figure 1** ELR (Ethics, Lawfulness and Robustness) Model

**2.3.2 AI Development Lifecycle.** The AI Development Lifecycle is represented in Figure 2, which is another model borne out of this research and presented to the standards TC62 for review. This AIDL is intended to convey to AIeMD developers the basic components of AI and how these might fit into the traditional SDLC. A draft version of this diagram was also reviewed and accepted by experts of TC62/MT49 for inclusion in the IEC 62304 ED2 standard anticipated for release in 2027/2028, with the final updated version presented here for publication in advance of inclusion in the standard.



**Figure 2** AI Development Lifecycle Model

The components of AIDL include *AI Planning & Design*, *Data Engineering* *and Management*, *Model Building* and *Tuning*, *Verification* and *Validation*, *Deployment* and finally *Model Maintenance* and *Decommission* which includes continuous monitoring. These components must ensure an appropriately defined AI Data Management Lifecycle that feeds into the AI Development lifecycle, as shown. Although some of these components are not new, the AI relationship and integration of AI into the existing SDLC is new and is important for adequate understanding and realising trustworthiness objectives. These have not been defined for AIeMD development or been specified as a holistic set of requirements. This research presents the State of the Art on aspects necessary for AIeMD development, that industry is ‘*crying out’* for, according to expert review analysis performed under this research. Also, regulators require a consistent yard stick with which to assess AIeMDs for release to market. Once deployed, an adequate Change Management and Monitoring system is essential and proposed in **Figure 3**.



**Figure 3** AI Change Management (AICM) Model for TwAI in Healthcare

As AI standards for healthcare are under development, requirements are not yet sufficiently detailed or complete to support AIeMD software development and validation [27], [29], [30], [31]. This has been confirmed by this research through TC, standards development discussions and expert interviews of AI developers from leading organisations. Hence, this work is aimed at supporting the development of these standards to ensure sufficient detail is provided, including the provision of the AIDL, ELR and AICM models. These models together can be taken as an TwAI framework for healthcare that will meet regulatory requirements and are necessary to meet the medical device requirements and those of the EU AI Act [32].

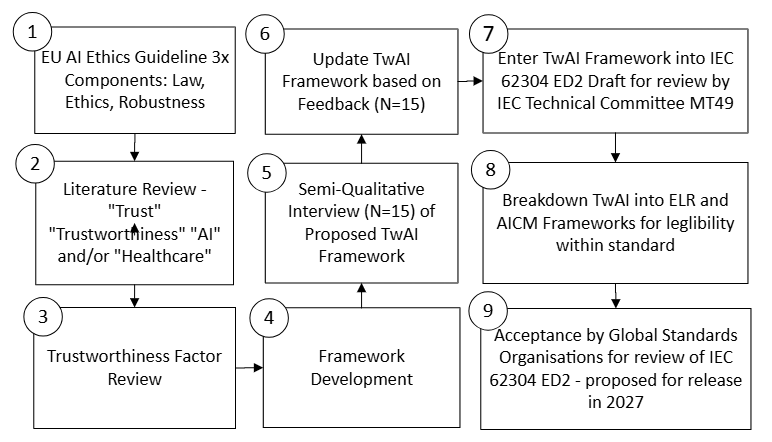
**2.3.3 Moving beyond the Conceptual Level.** This research moves beyond the conceptual level of *trust* within healthcare and identifies a more prescriptive level of TwAI and its sub-parts. The European Commission (EC) Ethics Guideline identified *Trustworthy AI* (TwAI) as their *‘foundational ambition’* [7]*.* In support of this, they define three primary components necessary for trustworthy AI: 1) *Lawful*, 2) *Ethical,* and 3) *Robust* [7]. We utilise these in our framework as a premise and combine academic research with existing regulatory medical device requirements for identification of the necessary *verifiable* measures necessary to deliver TwAI within healthcare. The novelty is in the identification of relevant ethical values and integration of key components and concepts together to form part of AIeMD requirements that have not been seen before. Thus, providing a consistent measure for regulatory scrutiny that was called for by [19].

# 3 Framework Development

Both primary and secondary research was performed in the development of the framework presented in this research. This was initiated by a literature review related to TwAI, as presented in the methodology section. Although recent scholars have discussed the epistemological and ethical basis of trust within AI, none have gone so far as to propose a structured formulae for achieving a consistent measure capable of delivering TwAI within healthcare. We can see from our review of current research that TwAI has not yet been sufficiently defined with various approaches and understanding across researchers, presented in the Appendix.

**3.1 Methodology**

The methodology used in the creation of the proposed framework as part of this research is presented in **Figure 4** Methodology for Creating Framework.



**Figure 4** Methodology for Creating Framework

A literature review and analysis of existing medical device regulatory requirements with international standards was utilised to create the original ELR and AICM models [31]. The ELR and AICM were presented as one TwAI Framework to participants subject to interview under this research. This led to update and improvement based on expert feedback. For clarity and legibility, it was decided to split the ELR and AICM into separate diagrams, which can be considered separately or taken together and used within the AIDL to represent a TwAI Framework for healthcare of AIeMD. The final framework is built upon the existing medical device (MD) regulatory landscape, with concepts familiar to manufacturers to ensure inclusion within their existing Quality Management System (QMS) is feasible and practicable.

**3.1.1 Design.** The literature review is used for generation of a meta-synthesis of mixed, qualitative, and quantitative studies together with a review of current regulatory requirements and standards within medical device development for use in healthcare. The factors identified as key to TwAI within these studies were qualitatively reviewed by experts for completeness. Validation of the framework presented here was performed by expert reviews through standards TC and semi-qualitative interview by experts in industry as described here.

**3.1.2 Method.** A systematic literature review and qualitative design approach was employed.

*3.1.2.1 Literature Review:*A search was conducted on the following databases to identify studies published in English: Academic Search Complete (EBSCOhost) and MEDLINE, accessed up to 06 Feb 2024. The meta-aggregation approach is based on the inclusion criteria and appropriate assessment of research methodologies were used to summarise findings from qualitative analysis focusing on definitions of TwAI within healthcare. EU Commission and FDA medical device databases and guidance documents ([MDCG Guidance](https://health.ec.europa.eu/medical-devices-sector/new-regulations/guidance-mdcg-endorsed-documents-and-other-guidance_en), [International Standards Organisation (ISO)](https://www.iso.org/home.html), [FDA Recognised Consensus Standards](https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfStandards/results.cfm) were also reviewed for current standards and guidelines related to AI in healthcare.

*3.1.2.2 Literature Review Inclusion and Exclusion Criteria:*The following were the inclusion criteria for the study.

1. **Search terms:** Artificial Intelligence or A.I. or AI + healthcare or medical + trustworthy or trust or trustworthiness.
2. **AI application:** medical device, healthcare, clinical health, or wellbeing.
3. **Language**: English
4. **Design**: meta-analysis of literature review.
5. **Study types**: qualitative, mixed-methods or quantitative design approach; case studies and opinion papers were included. Peer reviewed research papers only.
6. **Publication Date:** 2008-2024.

*3.1.2.3 Literature Review Results:*Of the 3,908 studies screened, 43 met the inclusion criteria and incorporated into analysis. A total of 30 papers were utilised to synthesise findings and used as the basis of the conclusions drawn here. A summary of results and citations are provided under the Appendix and supports the rationale for 30 papers included.

*3.1.3.1 Semi-Qualitative Interview:* SQI’s were conducted to identify the needs of AI developers within healthcare and their current understanding of the regulations and requirements for AIeMD development. Expert review of the TwAI framework (ELR and AICM together) were assessed during Semi-Qualitative Interview (n=14) and one expert review over email (n=1) was performed. Hence expert review of the framework was performed by fifteen experts (n=15[[1]](#footnote-2)) within twelve (12) different AI development organisations within healthcare. These spanned various geographic locations including the US, Europe (France, Germany, Ireland, The Netherlands), Canada, UK, Scotland, Switzerland, and Pakistan. The AIDL together with the ELR and AICM was subsequently reviewed by TC62 expert reviews and accepted initially by standards organisations under Committee Draft issued internationally for comment and feedback.

*3.1.3.2 Inclusion / Exclusion Criteria for Expert Interview:*Only organisations from healthcare industry actively developing AIeMD were included in the study.

*3.1.3.3 Interview Results Summary:* Analysis of the expert opinion and interview data identified a lack of understanding of requirements for AIeMD development. All indicated continuing to follow IEC 62304 ED1 which has no reference to AI whatsoever. Only one (1) interviewee affirmed that they had updated their QMS and existing SDLC with AI related requirements, based on German Notified Body Guidance. All others indicated that they had not yet updated their procedures where regulatory requirements and AI healthcare standards were not yet available. All indicated a need for clarification of requirements, and all indicated a need for understanding as to what should be included for ‘ethical values’ or ‘fundamental rights’ called out specifically by the EU AI Act. There was a lack of understanding of what measures to include, apart from ‘accuracy’ that was mentioned in most cases when asked what performance measures were being used for performance claims. There was a further lack of understanding as to the various types of ‘accuracy’ and the ways in which overfitting can impact results. When asked about post-market monitoring, all indicated the existing post-market surveillance plan would perform monitoring in the typical manner, which when probed amounted to feedback and complaint handling for the most part. It was evident from the analysis of the interviews that the QMS had not been updated by the majority of organisations beyond the traditional model that is not fit for purpose or suitable to deliver TwAI in healthcare.

**3.1.4 Research Conclusion.** The results of this research demonstrate that there is inconsistent terminology, definitions and assumptions underlying TwAI, particularly in healthcare. This research found a lack of well-defined processes, methodologies and standards to ensure developers of AIeMD apply consistent and comparable practices to development and monitoring of clinical performance outputs. This is a challenge for both developers and regulators. The risk is that this lack of requirements and consistent application may be more keenly felt in the post-market phase following AIeMD release to the market. Application of a robust AI Management System, AI Development and Data Management Lifecycle is necessary. The frameworks presented here, when consistently applied, can provide TwAI that both developers and regulators can use to meet regulations and stakeholder needs. Ongoing monitoring in the post-market phase is necessary to ensure continued safety and performance where AI is prone to autonomous change. Oversight and continuous monitoring to a defined set of measurable performance claims is necessary for both locked and adaptive AI algorithms [33]. Without all these aspects, TwAI cannot be successfully achieved in healthcare.

**3.1.5 Limitations.** Not all literature was available for review given access or language constraints (English) and therefore, the results may be limited or biased to available resources or English-speaking territories. The JBI Qualitative Critical Appraisal Checklist was not used to evaluate the quality of the included studies given the variation in research methodology listed in the appendices. Instead, literature screening process and results adopted the PRISMA workflow method also presented in the Appendix. Geographic bias may also impact on the expert review feedback with different regulatory perspectives. The approach taken is limited to the EU only. The TwAI Framework is a starting point amidst an evolving regulatory landscape and will require updating by organisations over-time. Furthermore, the framework has yet to be implemented and validated in real-world AIeMD development project through industry adoption. The future aim is to develop a practical tool for Change Management of AIeMDs that incorporate the necessary elements of the TwAI Framework described here. Hence, providing a measurement of trustworthiness in practice.

**4** **Trustworthy AI Framework in Healthcare**

This research contributes to the State of the Art by providing a TwAI framework of three interconnected models for use by medical device manufacturers and developers involved in AIeMD development. Additionally, it can be used by regulators and standards bodies for inclusion in healthcare specific AI guidance and standards. The TwAI Framework proposed here is in three parts.

**Figure 1** ELR (Ethics, Lawfulness and Robustness) model is presented as the top-level of the TwAI hierarchy for healthcare. We have identified appropriate ethical values required to meet the fundamental rights necessary to meet the EU AI Act. This is new to medical devices manufacturers, where the focus has not been so explicit on the rights of citizens, but instead addresses safety and performance of the device itself. We provide clear identification of existing ethical values and fundamental rights arising from relevant EU agreements, that are intended to underpin our laws and are accepted within the EU. The ELR Framework identifies the 3-key pillars from the AI Ethics Guideline [7] *Lawfulness, Ethics* and *Robustness* and is extended to explain the components to be included in an SDLC or AI Development Lifecycle.

**Figure 2** AI Development Lifecycle Model is presented next in which we provide a model for adoption by manufacturers that will fit into their existing QMS. Without this work, regulators may not know which aspects to assess for AI related devices creating significant challenge for both regulators and manufacturers. Also, without an adequate framework as presented here, developers are free to define their own. However, we have seen from our research that manufacturers under scrutiny by regulators, would rather wait and rely on the existing traditional standards with no mention of AI (IEC 62304), rather than risk updating their QMS without appropriate guidance from published literature or standards. This is in a large part from fear of regulatory bodies issuing non-conformance. This only creates more work for manufacturers. Whereas, to do nothing and continue innovation without constraint, takes the least burdensome approach, though questionable results in terms of the assurance of AIeMDs under development. This also has the potential for greater level of defects or false positives/false negatives in the field. We saw this clearly from our interviewee based in Germany, who had some guidance provided by IG-NB [34]. This provided the organisation with the necessary guidance that they can rely on in audit and claim current state of the art, as mandated by MDR 2017/745.

**Figure 3** AI Change Management (AICM) Model for TwAI in Healthcare extends the TwAI framework still further. We have broken down *Robustness*, which was a key pillar of trustworthiness in the EU Ethical Guideline already discussed. We identify what robustness means to an AIeMD developer in operational terms. This AICM model provides key characteristics necessary to meet both the MDR 2017/745 and EU AI Act that are necessary for safety, reliability, security and performance, amongst other requirements (usability, explainability, transparency, etc.). Key to this framework is the addition of the premarket AI performance measures, such as accuracy, mean accuracy, sensitivity, specificity, precision, ROC, F1, Error Score and more. It is noted in the framework that an adequate rationale should be provided for selection of the specific performance metrics relevant to the algorithm and device type. This is key to the strength of TwAI, where existing models typically promote a measure of accuracy without understanding how this might change over time or how it is monitored. It is necessary to understand and document the underlying assumptions beyond simply pointing to one measure of accuracy at a given point in time, that does not provide the full story of an AI algorithm in clinical use. Hence, the AICM identifies various measures relevant to AI that must all be assessed for applicability early in development and in the post-market phase, which can then provide the level of regulatory compliance and transparency needed [11].

Transparency focuses on the requirements from a stakeholder perspective and performance focuses on the verifiable measurement attributes that are necessary for demonstration of compliance to manufacturer claims. These must be made known to regulators, clinicians and patients. We have seen from the literature review and expert reviews that ambiguity surrounds TwAI requirements for AIeMD. With this in mind, how are requirements to be adequately defined within an AIDL without this research? We know from the literature that not identifying adequate requirements early in the SDLC process is a primary reason for defects in software medical devices [35]. We cannot afford to continue in this manner, particularly with AI.

This research aims to provide a comprehensive framework that provides the necessary building blocks for safe and reliable AIeMD. The performance characteristics identified in the proposed AICM model are necessary requirements that must be subject to frequent routine checks and oversight monitoring. These act as an input into the post-market surveillance phase and change management processes, supporting safety, performance and transparency through an iterative closed-loop cycle.

We have drawn on the current regulatory landscape to act as the backdrop to the newly identified AI-related frameworks and performance attributes relevant to AIeMD safety and performance. This ensures that developers of AIeMD can integrate these frameworks into their existing processes. This ensures innovation can continue without undue burden and cost, whilst providing TwAI in Healthcare.

**5 Discussion**

Although some AI standards are under development, none have yet provided a comprehensive TwAI Framework for inclusion within an AI Management System for development of safe, reliable, secure and effective AIeMD. Furthermore, none have provided a sufficiently comprehensive and detailed framework that identifies verifiable measurement attributes that are necessary in determining robust safety and performance characteristics specific to AIeMD. The frameworks presented in this research close the gap for AIeMD developers and regulators by providing a comprehensive TwAI model capable of meeting the EU AI Act. The intention is to demonstrate through use in healthcare that application of these frameworks together will yield safe and reliable devices with improved results for patients and clinicians.

This research demonstrates the lack of consensus as to what is required to achieve TwAI within healthcare and hence, we propose a harmonised approach based on these results. **Figure 1**, **2** and **Figure 3** together form a proposed TwAI Framework that is holistic and comprehensive for AIeMD development within healthcare. This supports the existing regulatory landscape for medical devices. Although other frameworks are beginning to emerge in the literature, none to date cover the scope of AIeMD sufficiently or with the required degree of completeness for operationalisation in healthcare. Nor do they have the appropriate level of transferable terminology necessary for understanding by legal manufacturers to upgrade their QMS. One example is the model presented by Solanki and colleagues [36] that is somewhat similar our top-level ELR structure though without providing specifics for inclusion in the requirements as we have. The proposed framework here differs quite significantly at the operational level, where medical device design control, transparency, and performance requirements for development of AIeMD is presented in terms that can be understood by developers and regulatory personnel employed within the medical device industry. We have ensured suitable and adequate performance variables are identified that can be operationalised within an existing QMS and SDLC that needs to make accommodations for AIeMDs.

We have taken a top-down approach by starting with the ELR and moving through the AIDL and then through to the AICM to build a complete framework. This provides the necessary clarity to AIeMD developers and manufacturers, many of whom have indicated through this research that they are not aware of the relevant ethical values that exist or are applicable to AIeMD development. Most developers simply make reference to bias and trust as generic concepts to be considered under ethics for AI. Moreover, through external validation of this research from expert reviews, those interviewed expressed support and even some ‘*delight’* at such guidance emerging.

**5.1 Delivering TwAI in Healthcare**

The achievement of TwAI in healthcare requires both AIeMD development and clinical performance to have a common understanding of the requirements from ethics and fundaments rights of patients, through to delivery of devices on the market. Trustworthy AIeMD can only come about through delivery of positive interactions with devices in use. Patients will have varying degrees of *trust* in a test result based on continual evidential scenarios that occur in their life and their families’ lives. We do not ask them to simply *trust* the result because we may or may not fully understand the mechanism of action of an AI algorithm. The performance of the test or treatment must be reviewed against defined requirements and performance standards. Instead, we build trust through proven methods of delivering quality, safety, reliability, security and performance.

In short, patient and clinician trust are earned through positive affects with AIeMDs that require appropriate and consistent regulatory review of the performance characteristics against claims cited. These claims are necessary for regulatory authorisation to allow placement on the market and for on-going monitoring. Once released to the healthcare facility or home use, further clear responsibilities are necessary to be defined with continued monitoring and oversight by relevant stakeholders with full transparency. The framework presented demonstrates how the continuous monitoring of performance of the AI system allows for a closed-loop system with transparency to ensure the appropriate level of communication, awareness and explainability suitable to deliver TwAI in healthcare.

**6 Conclusion**

The research presented here demonstrates the varied parameters and constructs that make-up TwAI. This research identifies the need for a harmonised approach and extends the current literature to suggest that we must go beyond theoretical and psychological constructs of ethical values. Although our ethical values are already defined, they are not currently or explicitly applied to development of devices within healthcare. This is particularly true of the SDLC and AIeMD development processes, where identification of risks to fairness or bias is given some consideration. Little is known by AIeMD developers of other applicable ethical values that are intended to be represented and assessed. If manufacturers and AI developers are to be expected to consistently and methodically operationalise appropriate measures capable of yielding TwAI in healthcare, a common understanding of what is necessary must made publicly available. To that end, we identify a range of processes and parameters that are believed to underpin TwAI specific to healthcare. This research proposes a TwAI framework that is cognizant of the regulatory landscape within which AIeMD developers work.

The TwAI framework presented by this research is a starting point for operationalising TwAI in Healthcare. The aim is to identify key characteristics and performance measures that must be fully assessed within the AI development life cycle and throughout the post-market surveillance phase, ensuring ethical, legal, and robust performance of AIeMDs. Future work is underway to identify methods for change management with monitoring and oversight in the post-market phase, ensuring real-time notification of changes to AIeMDs that are necessary to assure safety and performance is delivered and maintained.

**Acknowledgements**. This research is supported by the Research Ireland (RI), Frontiers for the Future Programme under the Regulatory Compliance Framework for Trustworthy AI Medical Device Software (Reg-Fr-AIMs) project, ID 21/FFP-A/9255; a collaborative project undertaken by DkIT and UCD, Ireland.

# References

[1] K. Bærøe, A. Miyata-Sturm, and E. Henden, ‘How to achieve trustworthy artificial intelligence for health’, *Bull World Health Organ*, vol. 98, no. 4, pp. 257–262, Apr. 2020, doi: 10.2471/BLT.19.237289.

[2] I. Pacal and S. Kılıcarslan, ‘Deep learning-based approaches for robust classification of cervical cancer’, *Neural Comput Appl*, vol. 35, no. 25, pp. 18813–18828, Sep. 2023, doi: 10.1007/s00521-023-08757-w.

[3] B. Koopman, G. Zuccon, A. Nguyen, A. Bergheim, and N. Grayson, ‘Extracting cancer mortality statistics from death certificates: A hybrid machine learning and rule-based approach for common and rare cancers’, *Artif Intell Med*, vol. 89, pp. 1–9, Jul. 2018, doi: 10.1016/j.artmed.2018.04.011.

[4] C. Jones, J. Thornton, and J. C. Wyatt, ‘Artificial intelligence and clinical decision support: Clinicians’ perspectives on trust, trustworthiness, and liability’, in *Medical Law Review*, Oxford University Press, 2023, pp. 501–520. doi: 10.1093/medlaw/fwad013.

[5] IEC SC62A, ‘Design Specification for the second edition of IEC 62304’, Geneva, 2024.

[6] EN-62304, ‘Medical device software: software life-cycle processes’, Nov. 30, 2015, *CENELEC, Geneva*.

[7] EC AI HLEG, ‘Ethics Guidelines for Trustworthy AI’, Brussels, Apr. 2019. [Online]. Available: https://ec.europa.eu/digital-

[8] B. Li *et al.*, ‘Trustworthy AI: From Principles to Practices’, Jan. 13, 2023, *Association for Computing Machinery*. doi: 10.1145/3555803.

[9] R. Dlugatch, A. Georgieva, and A. Kerasidou, ‘Trustworthy artificial intelligence and ethical design: public perceptions of trustworthiness of an AI-based decision-support tool in the context of intrapartum care’, *BMC Med Ethics*, vol. 24, no. 1, Dec. 2023, doi: 10.1186/s12910-023-00917-w.

[10] V. Laflèche and D. Sturtewagen, ‘CEN-CLC Strategy 2O3O’, Jan. 2021. Accessed: Apr. 17, 2025. [Online]. Available: CEN -

[11] MDCG-2025-6, ‘Medical Devices Interplay between the MDR & IVDR and the Artificial Intelligence Act (AIA)’, Brussels, Jun. 2025. Accessed: Jun. 20, 2025. [Online]. Available: https://health.ec.europa.eu/document/download/b78a17d7-e3cd-4943-851d-e02a2f22bbb4\_en?filename=mdcg\_2025-6\_en.pdf

[12] P. Thagard, ‘Societies of minds: Science as Distributed Computing’, © Paul Thagard, 1993.

[13] F. Flores and R. C. Solomon, ‘Creating Trust’, 1998. [Online]. Available: https://www.jstor.org/stable/3857326

[14] J. A. Simpson, ‘Psychological Foundations of Trust’, 2007.

[15] V. Gligor and J. M. Wing, ‘Towards a Theory of Trust in Networks of Humans and Computers’, *Security Protocols LNCS*, pp. 223–243, 2011.

[16] C. Castelfranchi and R. Falcone, *Trust Theory: A Socio-Cognitive and Computational Model*. UK: John Wiley & Sons, 2010.

[17] J. Ma and M. A. Orgun, ‘Trust management and trust theory revision’, *IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans*, vol. 36, no. 3, pp. 451–460, May 2006, doi: 10.1109/TSMCA.2006.871628.

[18] M. B. Unver, ‘Governing fiduciary relationships or building up a governance model for trust in AI? Review of healthcare as a socio-technical system’, *International Review of Law, Computers and Technology*, vol. 37, no. 2, pp. 198–226, 2023, doi: 10.1080/13600869.2023.2192569.

[19] G. Starke, R. van den Brule, B. S. Elger, and P. Haselager, ‘Intentional machines: A defence of trust in medical artificial intelligence’, *Bioethics Wiley*, vol. 36, pp. 154–161, Apr. 2020.

[20] N. Omrani, G. Rivieccio, U. Fiore, F. Schiavone, and S. G. Agreda, ‘To trust or not to trust? An assessment of trust in AI-based systems: Concerns, ethics and contexts’, *Technol Forecast Soc Change*, vol. 181, Aug. 2022, doi: 10.1016/j.techfore.2022.121763.

[21] R. C. Feldman, E. Aldana, and K. Stein, ‘Artificial Intelligence in the Health Care Space: How We Can Trust What We Cannot Know’, Center for Innovation, University of California Hastings College of the Law, Jul. 2019. [Online]. Available: https://www.ibm.com/ibm/history/ibm

[22] J. Zhang and Z. ming Zhang, ‘Ethics and governance of trustworthy medical artificial intelligence’, *BMC Med Inform Decis Mak*, vol. 23, no. 1, Dec. 2023, doi: 10.1186/s12911-023-02103-9.

[23] WMA, ‘Declaration of Helsinki-Ethical Principles for Medical Research involving Human Subjects’, Oct. 2023.

[24] EU, *MDR 2017/745*. 2017, pp. 1–336. Accessed: Jun. 19, 2025. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2017/745/oj/eng

[25] ISO-14155, ‘Clinical investigation of medical devices for human subjects-good clinical practice’, 2020.

[26] ISO-20916, ‘In vitro diagnostic medical devices - performance studies using specimens from human subjects-good study practice’, 2019. [Online]. Available: www.iso.org

[27] E. B. Turpin, Hoefer AAMI, J. A. Lewelling, P. P. Baird, and et al., ‘Machine Learning AI in Medical Devices: adapting regulatory frameworks and standards to ensure safety and performance’, 2020. [Online]. Available: www.aami.org

[28] B. Pickering, ‘Trust, but verify: Informed consent, ai technologies, and public health emergencies’, *Future Internet*, vol. 13, no. 5, May 2021, doi: 10.3390/fi13050132.

[29] A. Arora *et al.*, ‘The value of standards for health datasets in artificial intelligence-based applications’, *Nat Med*, Nov. 2023, doi: 10.1038/s41591-023-02608-w.

[30] A. A. H. de Hond *et al.*, ‘Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review’, Dec. 01, 2022, *Nature Research*. doi: 10.1038/s41746-021-00549-7.

[31] N. St John Lynch, R. Loughran, M. McHugh, and F. McCaffrey, ‘Artificial Intelligence-Enabled Medical Device Standards: A Multidisciplinary Literature Review’, G. Yilmaz, P. Clarke, A. Riel, R. Messnarz, C. Greiner, and T. Peisl, Eds., Munich: Springer, Cham., Sep. 2024, pp. 112–130. doi: 10.1007/978-3-031-71139-8\_8.

[32] EU, *The AI Act 2024/1689*. Brussels, 2024, pp. 1–144. Accessed: Apr. 08, 2024. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L\_202401689

[33] P. Dhunnoo, ‘Locked and Adaptive Algorithms in Healthcare Differences, Importance and Regulatory Hurdles - The Medical Futurist’, *Artificial Intelligence in Medine*, vol. Online, Feb. 2022.

[34] IG-NB and Team-NB AI Task Force, ‘Questionnaire on Artificial Intelligence in Medical Devices’, Dec. 2024. [Online]. Available: https://www.team-nb.org/team-nb-documents/

[35] M. McHugh, A. Abder-Rahman, and F. McCaffery, ‘The Significance of Requirements in Medical Device Software Development’, in *EuroSPI*, Nov. 2013. Accessed: Feb. 24, 2024. [Online]. Available: https://eprints.dkit.ie/id/eprint/311

[36] P. Solanki, J. Grundy, and W. Hussain, ‘Operationalising ethics in artificial intelligence for healthcare: a framework for AI developers’, *AI and Ethics*, vol. 3, no. 1, pp. 223–240, Feb. 2023, doi: 10.1007/s43681-022-00195-z.

[37] P. Tugwell and D. Tovey, ‘PRISMA 2020 - Editor’s Choice’, *J Clin Epidemiol*, vol. 134, pp. A5–A6, 2021.

[38] M. J. Page, J. E. McKenzie, P. M. Bossuyt, and et al., ‘The PRISMA 2020 Statement: an updated guideline for reporting systematic reviews’, *J Clin Epidemiol*, vol. 134, pp. 178–189, 2021.

[39] S. Ali *et al.*, ‘Metaverse in Healthcare Integrated with Explainable AI and Blockchain: Enabling Immersiveness, Ensuring Trust, and Providing Patient Data Security’, *Sensors*, vol. 23, no. 2, Jan. 2023, doi: 10.3390/s23020565.

[40] H. Bangui, B. Buhnova, and M. Ge, ‘Social Internet of Things: Ethical AI Principles in Trust Management’, in *Procedia Computer Science*, Elsevier B.V., 2023, pp. 553–560. doi: 10.1016/j.procs.2023.03.070.

[41] P. Dhiman *et al.*, ‘Healthcare Trust Evolution with Explainable Artificial Intelligence: Bibliometric Analysis’, Oct. 01, 2023, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/info14100541.

[42] N. Díaz-Rodríguez, J. Del Ser, M. Coeckelbergh, M. López de Prado, E. Herrera-Viedma, and F. Herrera, ‘Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation’, *Information Fusion*, vol. 99, Nov. 2023, doi: 10.1016/j.inffus.2023.101896.

[43] A. Dy *et al.*, ‘AI improves accuracy, agreement and efficiency of pathologists for Ki67 assessments in breast cancer’, *Sci Rep*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-51723-2.

[44] L. Escudero Sanchez *et al.*, ‘Integrating Artificial Intelligence Tools in the Clinical Research Setting: The Ovarian Cancer Use Case’, *Diagnostics*, vol. 13, no. 17, Sep. 2023, doi: 10.3390/diagnostics13172813.

[45] F. Li, N. Ruijs, and Y. Lu, ‘Ethics & AI: A Systematic Review on Ethical Concerns and Related Strategies for Designing with AI in Healthcare’, Mar. 01, 2023, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/ai4010003.

[46] J. Fehr *et al.*, ‘Piloting a Survey-Based Assessment of Transparency and Trustworthiness with Three Medical AI Tools’, *Healthcare (Switzerland)*, vol. 10, no. 10, Oct. 2022, doi: 10.3390/healthcare10101923.

[47] T. Hagendorff, ‘The Ethics of AI Ethics: An Evaluation of Guidelines’, *Minds Mach (Dordr)*, vol. 30, no. 1, pp. 99–120, Mar. 2020, doi: 10.1007/s11023-020-09517-8.

[48] S. Kerstan, N. Bienefeld, and G. Grote, ‘Choosing human over AI doctors? How comparative trust associations and knowledge relate to risk and benefit perceptions of AI in healthcare’, *Risk Analysis*, 2023, doi: 10.1111/risa.14216.

[49] W. Liu *et al.*, ‘Explainable AI for Medical Image Analysis in Medical Cyber-Physical Systems: Enhancing Transparency and Trustworthiness of IoMT’, *IEEE J Biomed Health Inform*, 2023, doi: 10.1109/JBHI.2023.3336721.

[50] OECD, ‘Advancing Accountability in AI - Governing and managing risks throughout the lifecycle for Trustworthy AI’, 2023. [Online]. Available: www.oecd.ai/wips.

[51] S. Sachan, F. Almaghrabi, J. B. Yang, and D. L. Xu, ‘Evidential reasoning for preprocessing uncertain categorical data for trustworthy decisions: An application on healthcare and finance’, *Expert Syst Appl*, vol. 185, Dec. 2021, doi: 10.1016/j.eswa.2021.115597.

[52] S. Saini and N. Saxena, ‘A Survey of Threats to Research Literature-dependent Medical AI Solutions’, *ACM Comput Surv*, vol. 55, no. 14 S, Jul. 2023, doi: 10.1145/3592597.

[53] J. L. Schultze, ‘Building Trust in Medical Use of Artificial Intelligence – The Swarm Learning Principle’, *Journal of CME*, vol. 12, no. 1, Dec. 2023, doi: 10.1080/28338073.2022.2162202.

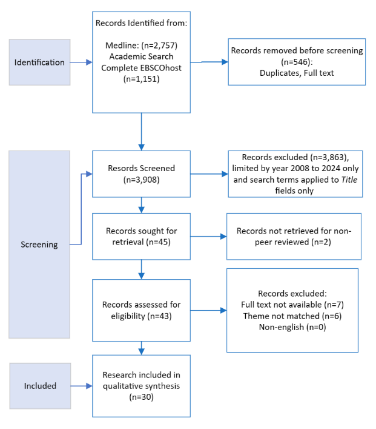
[54] P. J. Schulz, M. O. Lwin, K. M. Kee, W. W. B. Goh, T. Y. T. Lam, and J. J. Y. Sung, ‘Modeling the influence of attitudes, trust, and beliefs on endoscopists’ acceptance of artificial intelligence applications in medical practice’, *Front Public Health*, vol. 11, 2023, doi: 10.3389/fpubh.2023.1301563.

[55] M. Simion and C. Kelp, ‘Trustworthy artificial intelligence’, *Asian Journal of Philosophy*, vol. 2, no. 1, Mar. 2023, doi: 10.1007/s44204-023-00063-5.

[56] S. Thiebes, S. Lins, and A. Sunyaev, ‘Trustworthy artificial intelligence’, *Electronic Markets*, vol. 31, no. 2, pp. 447–464, Jun. 2021, doi: 10.1007/s12525-020-00441-4.

[57] C. Wu, H. Xu, D. Bai, X. Chen, J. Gao, and X. Jiang, ‘Public perceptions on the application of artificial intelligence in healthcare: a qualitative meta-synthesis’, *BMJ Open*, vol. 13, no. 1, Jan. 2023, doi: 10.1136/bmjopen-2022-066322.

[58] P. Xiong, S. Buffett, S. Iqbal, P. Lamontagne, M. Mamun, and H. Molyneaux, ‘Towards a robust and trustworthy machine learning system development: An engineering perspective’, *Journal of Information Security and Applications*, vol. 65, Mar. 2022, doi: 10.1016/j.jisa.2022.103121.

**Appendix**

**Figure 5** is provided as an overview of the literature screening process and results, with adoption of the PRISMA workflow method [37], [38].

This is followed by **Table 1.** With citations (n=30) research papers reviewed for an understanding of trustworthiness.

**Figure 5** PRISMA Literature Screening Process & Results

**Table 1**. Literature Review Results

|  |  |
| --- | --- |
| Study | Main Results |
| 1. Ali et al. (2023) | Safety, Security, Privacy of Data, Trust, explainability, interpretability and transparency, traceability, authentication, authorisation, audit logs, accountability. |
| 1. B. Li et al. (2023) | Fairness, privacy, accountability, generalisation, transparency, robustness, reproducibility, auditability, safety, autonomy, sustainability, performance, security |
| 1. Bærøe et al. (2020) | Protection, Mistrust of Suggests that globalised efforts are necessary to safeguard against potential harms of AI in healthcare. |
| 1. Bangui et al. (2023) | Transparency, Justice & Fairness, Non-maleficence, Responsibility, Privacy, Freedom & Autonomy, Trust, Sustainability, Dignity, Solidarity (cohesion). |
| 1. Dhiman et al. (2023) | Interpretability, Transparency, Explainability**:** explainable models, opaque models, transparent models, under/over-reliance. |
| 1. Díaz-Rodríguez et al. (2023) | Confirms 7 requirements from EC Ethics Guideline and diverging view of TwAI concluding, TwAI and Responsible AI systems are crucial for future of our society. |
| 1. Dlugatch et al. (2023) | Fairness and reliability as foundation of Trust. Public institutions are trusted over private companies. Not so interested in list of design features but how AI might undermine or promote ethical values. |
| 1. Dy et al. (2024) | Inter-rater variability, Index Error, Reliability, Accuracy, Consistency of scoring[results], noted: “83% trust it will improve efficiency”, clinical agreement with significance statistically assessed. AI assisted scoring was superior in accuracy and reduced error across all pathologist subdisciplines. |
| 1. EC AI HLEG (2019) | Trustworthy AI has 3 components: lawful, ethical, and robust. 4 ethical principles (foundations): respect for autonomy, prevention of harm, fairness, and explicability. 7 Key requirements for realisation: human agency & oversight, technical robustness and safety, privacy and data governance, transparency, diversity, non-discrimination and fairness, societal and environmental wellbeing, accountability. Operationalise key requirements: “*tailor to specific AI application*” |
| 1. Escudero Sanchez et al. (2023) | Accuracy and reproducibility of predictions. Automatic detection and precisely delineating disease. |
| 1. F. Li et al. (2023) | 12-Main ethical issues: justice & fairness, freedom & autonomy, privacy, transparency, patient safety and cyber security, trust, beneficence, responsibility, solidarity, sustainability, dignity, and conflicts (with 19-sub issues, e.g., informed consent, confidentiality, explainability, etc.). |
| 1. Fehr et al. (2022) | AI ethics guideline remains unfulfilled. Close transparency gap for medical AI. Intended Use and Clinical considerations clear & transparent. Reporting gaps will limitation to trustworthiness. |
| 1. Feldman et al. (2019) | Trust and interpretability**,** explainability of black box  Societal Trust: *Provider competence* (certification from regulatory & standards), *patient interest*, and *information integrity*. |
| 1. Hagendorff (2020) | Reviews 22 major AI ethical guidelines and provides recommendations on overcoming their ineffectiveness. Accountability, privacy, and fairness account for 80%. |
| 1. Kerstan et al. (2023) | Human involvement in decision and treatment. Perceptions and preferences of humans must be addressed through transparency, AI literacy, explainability, multi-stakeholder engagement and co-creation (collaborative prototyping). |
| 1. Liu et al. (2023) | XAI can enhance transparency and trustworthiness in decision making. Improves accuracy of classification, removes low-quality images, and enables explainability of results. Improved interpretability and enhanced robustness. |
| 1. OECD (2023) download Jan 2024. | Fairness, Privacy & Data governance, Respect for Human Rights, Robustness & Digital Security, Transparency & Explainability & others listed under Tools for Trustworthy AI (i.e., safety, human wellbeing, sustainability) |
| 1. Omrani et al. (2022) | Unequal degrees of trust across EU28 (member states). TwAI focuses on discrimination, responsibility, and accountability, with different demographics having different opinions. 3 learnings; 1) strengthen technical and social robustness of AI systems, 2) create systems compliant with law and regulations, 3) guarantee compliance with ethical principles and values. |
| 1. Sachan et al. (2021) | Uncertainty, ambiguity & incompleteness of missing values at input and output, categorical attribute assignment incomplete and ambiguous. 3 types: informational uncertainty, unforeseeable uncertainty, uncertainly due to lack of pre-modelling explainability. |
| 1. Saini and Saxena (2023) | Security, integrity, credibility. |
| 1. Schultze (2023) | Data Protection & Security, Sustainable, environmentally protective, State of the Art Decision-Making, Accessible, Available, Interpretable, Robustness (sufficiently large data training sets), Appropriate algorithms, Fair, Transparent |
| 1. Schulz et al. (2023) | Trust and belief were found to fully mediate the effects of attitude on AI Acceptance by clinicians. |
| 1. Simion and Kelp (2023) | Safety, justice and explainability; a review of trustworthiness in AI |
| 1. Solanki et al. (2023) | Security\*, Self-Direction\*, Benevolence\*, Universalism\*, Non-maleficence, Freedom & Autonomy, Dignity, Privacy, Beneficence, Responsibility, Trust, Transparency (communication), Solidarity (cohesion), Justice & Fairness, Sustainability |
| 1. Starke et al. (2020) | Risks & Uncertainty; Reliability, Competence (validity, accuracy of predictions), Intentions (conflicts of interests, transparency, and systemic bias and explainability). Do not trust blindly but with clearly defined acceptable thresholds. |
| 1. Thiebes et al. (2021) | Fives foundational principals: beneficence, non-maleficence, autonomy, justice, and explicability. Additionally cited by others: competency, ability, integrity, functionality, helpfulness, reliability, predictability, performance, purpose, process. |
| 1. Unver (2023) | Transparency, reliability and fairness, adaptiveness, competence, responsibility, accountability and sustainable. |
| 1. Wu et al. (2023) | Ethical concerns requiring reasonable supervision of medical AI as key. |
| 1. Xiong et al. (2022) | Support for verifiable claims; agrees need to move beyond ethical principles for mechanisms that AI practitioners can adopt to make and verify claims about AI systems, as evidence of responsible behaviour. |
| 1. Zhang and Zhang (2023) | Technically safe and reliable & respects fundamental human rights and universal human values (ethics); Data quality, algorithmic bias and errors, opacity, safety and security, responsibility attribution. |

1. One expert agreed to review offline making the contributors amount of a total of fifteen (15). Hence, the total interviewed were fourteen (14). This expert review provided in-depth review and feedback on the TwAI Framework over email correspondence and did not want to be interviewed. [↑](#footnote-ref-2)