

Ethical Requirements for Healthcare AI Software Systems: Emerging Results from a Multi-Stakeholder Interview Study

Yutan Huang ✉ 

Monash University, Melbourne, Australia

Chetan Arora ✉ 

Monash University, Melbourne, Australia

Tanjila Kaniy ✉ 

Swinburne University of Technology, Melbourne, Australia

Anuradha Madugalla ✉ 

Deakin University, Melbourne, Australia

John Grundy ✉ 

Monash University, Melbourne, Australia

Abstract

Healthcare AI software systems raise ethical concerns that are often discussed as high-level principles but are difficult to translate into concrete software engineering requirements. This emerging-results paper reports preliminary findings from a multi-stakeholder interview study on how practitioners experience and prioritise ethical concerns in healthcare AI software systems. We conducted 17 semi-structured interviews with healthcare professionals (n=5), healthcare researchers (n=6), and software engineers (n=6). Using thematic synthesis, we identified three AI use contexts and four recurring challenge categories: model reliability and quality, data quality and processing, human oversight and skills gaps, and institutional and regulatory constraints. Privacy and safety emerged as baseline concerns across roles, while bias, accountability, and transparency varied depending on practitioners' responsibilities and use contexts. Based on these preliminary findings, we propose role-specific patterns as lightweight requirements artefacts to translate practitioner priorities into actionable software engineering practices across requirements elicitation, system design, testing, and auditing. We present these patterns as an initial artefact for community feedback and as a first step toward a broader approach to ethical requirements engineering (RE) for healthcare AI software systems. Future work will validate and refine the patterns using a larger, more balanced sample, evaluate their usefulness through practitioner workshops, and develop a reusable catalogue with guidance on traceability, verification, and auditing. Long-term, our goal is to support healthcare AI teams in moving from abstract ethical principles to testable requirements that can be integrated into existing development and governance workflows.

2012 ACM Subject Classification Software and its engineering → Requirements analysis

Keywords and phrases Healthcare AI software systems, ethical requirements, requirements engineering, empirical software engineering, thematic synthesis, practitioner interviews

Digital Object Identifier 10.4230/LIPIcs.ESEM.2026.72

Category Emerging Results, Vision & Reflection Track Paper

1 Introduction

Artificial intelligence is increasingly embedded into healthcare software systems, including predictive models, AI-assisted imaging, clinical documentation tools, patient-facing applications, robotic support systems, and newer large language model (LLM)-based applications [17, 19]. In *healthcare AI-based software systems* (henceforth, *HAISS*), the stakes are acute: software



© Yutan Huang, Chetan Arora, Tanjila Kaniy, Anuradha Madugalla, and John Grundy; licensed under Creative Commons License CC-BY 4.0

20th International Symposium on Empirical Software Engineering and Measurement (ESEM 2026).

Editors: Robert Feldt, Maria Paasivaara, Daniel Mendez, Stefan Wagner, and Marvin Muñoz Barón;

Article No. 72; pp. 72:1–72:14



Leibniz International Proceedings in Informatics

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

behaviour is tied to sensitive patient data, clinical decision-making, and organisational accountability. Failures can therefore cause harm rather than prevent it, including poor predictive accuracy, racial or demographic bias, unsafe recommendations, and increased clinician workload [4]. From a requirements engineering (RE) perspective, this turns ethics from a post hoc discussion into a concrete engineering problem: ethical concerns must be elicited, negotiated among stakeholders, specified as verifiable requirements, and checked throughout design, validation, deployment, and audit. Yet in practice, teams often do not treat ethics as requirements; privacy and safety concerns are identified late in development, as incidents or ad-hoc checks, rather than as elicited and verifiable non-functional requirements (NFRs) [22, 18]. Because these concerns arise at handoffs among engineers, clinicians, and researchers, we frame them as a cooperative RE problem. As AI systems become increasingly sophisticated, ethical dilemmas and RE challenges arise. Early work in AI ethics emphasised transparency, fairness, accountability, and privacy [14], prompting numerous ethical guidelines and frameworks [15]. In *HAISS*, AI is integrated into diagnostic imaging, digital health applications used by patients and staff, predictive analytics, and robotic surgery – settings where ethical concerns are acute due to patient sensitivity and direct impact on human lives [12]. Researchers have advocated multi-scale approaches to govern AI ethics in healthcare, operationalised through SE practices that span requirements to deployment [25]. Despite the growing number of ethical guidelines and frameworks, a significant gap remains as they rarely incorporate the perspectives of the software engineers who design and validate *HAISS*, nor those of the healthcare practitioners whose daily use shapes how principles are realised, and they seldom translate high-level principles into concrete SE activities [14, 5, 27]. Real-world, multi-stakeholder evidence is therefore needed to ground ethical principles in practice.

In this paper, we treat ethical priorities as ethical requirements spanning (i) quality requirements (NFRs), (ii) constraints (e.g., data residency, access controls), and (iii) assurance/process requirements (e.g., audits, monitoring, incident response), and we explore how to operationalise them via lightweight If-Then requirement patterns. To do so, we conducted 17 semi-structured interviews with software engineers, healthcare professionals, and healthcare researchers across three countries. Our study is guided by three research questions (RQs): **RQ1.** In what contexts and use cases is AI currently applied in healthcare? **RQ2.** What challenges and limitations do practitioners report when deploying AI in *HAISS*? **RQ3.** Which ethical considerations of *HAISS* do practitioners prioritise, and how?

Contributions. In this emerging results paper, we provide an initial empirically grounded artefact and make three contributions to industrial SE practice:

1. **Empirical analysis of role-specific AI use.** We characterise how software engineers, healthcare professionals and researchers incorporate AI in practice, and show how these distinct usage patterns imply different requirements, verification, and integration concerns.
2. **Role-specific challenges.** We identify four recurring challenge categories across three roles, and locate where each calls for new requirements, tests, or deployment safeguards.
3. **Ethical priorities → SE artefacts.** We map how practitioners prioritise five ethical dimensions and derive lightweight If-Then patterns that operationalise these priorities into actionable requirements.

Paper Structure. Section 2 reviews background on AI in healthcare and ethics. Section 3 describes our research methodology. Section 4 presents the results. Section 5 discusses our findings and implications. Section 6 reports threats to validity. Section 7 concludes with directions for future work.

2 Background / Related Work

Ethics of AI use in Healthcare. Prior work on healthcare AI ethics has identified recurring principles and risks, but it has also been less often explained how they should be operationalised as concrete SE requirements. Reviews of AI in healthcare highlight concerns around transparency, fairness, bias, privacy, patient safety, and accountability, while also noting that practitioner perspectives are underrepresented in the implementation of ethical principles [24, 13]. Karimian et al. identify key principles such as privacy, fairness, explicability, and human autonomy, but also show that tools and processes for testing and upholding ethical requirements remain under-specified [16]. Li et al. similarly map ethical issues and proposed strategies for healthcare AI, while emphasising the need to translate these strategies into practice in system development [20]. In our prior work, we organised ethical concerns of generative AI around five recurring dimensions: safety, privacy, transparency, bias/fairness, and accountability, and synthesised mitigation strategies including governance frameworks, human oversight, auditability, and privacy-preserving methods [14]. Building on this work, we use these five dimensions to elicit and compare practitioners' ethical priorities.

Practitioner Perspectives on AI in Healthcare. Practitioner studies show that healthcare professionals recognise both the opportunities and risks of AI adoption. Prior qualitative work has identified concerns about transparency, equity, accountability, data privacy, and the effect of AI on patient care and professional relationships [9, 23, 2]. These studies provide valuable insight into clinical and nursing perspectives, but the role of software engineers remains less visible. Existing work often treats software engineers as technical implementers rather than stakeholders who shape ethical outcomes through requirements, design, validation, deployment, and maintenance decisions [7]. Building on this gap, we incorporate the perspectives of healthcare professionals, healthcare researchers, and software engineers to examine how ethical concerns are discussed across roles in *HAISS*.

3 Methodology

We conducted a qualitative interview study across three stakeholder groups to investigate how practitioners experience, encounter, and prioritise ethical concerns in *HAISS*. We chose semistructured interviews because our goal was to surface reasoning specific to each role and varying with context, which structured surveys cannot capture. We analysed the resulting transcripts using the thematic synthesis process recommended for software engineering by Cruzes and Dybå, which proceeds in five steps [8].

Research Questions. The study is guided by three research questions (RQs).

RQ1: In what contexts and use cases is AI currently applied in healthcare? We examine real-world applications and the roles AI plays in patient care, diagnostics, and operations, including how AI is integrated into workflows, the motivations behind adoption, and the applications supported.

RQ2: What challenges and limitations do practitioners report when deploying AI in *HAISS*? We examine implementation and integration issues, ethical dilemmas, and operational constraints, comparing experiences across roles to understand factors that hinder adoption and areas for improvement.

RQ3: Which ethical considerations of *HAISS* do practitioners prioritise, and how?

We aim to understand which of the five ethical dimensions carried over from our prior synthesis [14] (privacy, safety, bias, accountability, and transparency) practitioners consider most important.

Participant recruitment and sampling. In our study, we aimed to recruit a diverse sample of participants representing three key groups: healthcare professionals, healthcare researchers and software engineers, as mentioned above. We advertised the study through a multi-channel strategy across professional and social platforms (LinkedIn, WeChat, and X) and recruited from IT companies, clinics, hospitals, and universities across China, Australia, and India. The first author conducted all interviews between February and July 2025, both via video conference and in person. The final sample comprised 17 participants (8 female, 9 male) with 1 to 24 years of professional experience. We continued recruitment until we reached this sample and observed thematic recurrence across the three roles. The study was approved by our institutional human research ethics committee (Project ID [45523]).

Interview Protocol. Before the main study we conducted three pilot interviews, one per role, and made minor refinements to the wording and the order of sections based on the feedback. Each interview lasted approximately 30 minutes and followed a protocol with three phases aligned to the research questions, preceded by an introduction and demographics questions. The first phase, *Practical Use Scenarios*, asked participants to describe in detail how they use AI in their clinical practice, research, or development work, and the benefits they have experienced. The second phase, *Current Challenges and Limitations*, asked participants about the limitations of failures they had encountered, the specific incidents involved, and whether and how these had been resolved. The third phase, *Ethical Considerations*, asked participants to rank the five ethical dimensions from 1 (highest priority), to 5 (lowest), having first been given brief definitions and examples of each, and to justify their ranking and suggest any solutions.

Analysis. We analysed the transcripts following the five steps of thematic synthesis recommended for software engineering by Cruzes and Dybå [8]: (1) reading all transcripts and noting recurring ideas; (2) segmenting each transcript into units aligned with the three research questions; (3) coding each segment with concise labels (for example, AIUseCase or TransparencyIssue); (4) merging overlapping codes into sub-themes; and (5) grouping sub-themes into the higher-level themes used to structure our results. For RQ3, we report the distribution of ranks assigned to each dimension and qualitatively analyse the accompanying justifications to explain how priorities differed across roles.

Trustworthiness of the analysis. The first author performed the primary coding of all 17 transcripts. Because a single coder introduces a risk of interpretation bias, we took two steps to make the analysis traceable. First, we maintained an audit trail: every code records the transcript identifier and the verbatim segment it was derived from, so each sub-theme and theme can be traced back to the utterances that produced it. Second, we attended explicitly to cases that diverged from the emerging pattern, most notably P12, who declined to produce a single ranking, and P16, whose account of accountability differed from that of every other software engineer.

■ **Table 1** Participant summary.

Group	n	Countries	Experience	AI usage
Healthcare professionals	5	China, Australia	3–24 yrs	Rarely–Always
Healthcare researchers	6	China, Australia	1–10 yrs	Sometimes–Always
Software engineers	6	India, Australia	2–10 yrs	Sometimes–Always
Total	17	China (8), Australia (8), India (1)	1–24 yrs	Rarely–Always

From Findings to Artefacts. To turn our findings into an actionable artefact, we derived role-specific If-Then patterns through a systematic mapping procedure. For each recurring challenge identified in RQ2, we identified the ethical dimension from RQ3 that it most directly threatens, and then formulated a concrete SE response that combines participants’ suggestions with the corresponding regulatory or technical obligation. Each resulting pattern specifies a trigger condition drawn from our interview data, a response, the responsible role, and the applicable guideline. This mapping is the methodological bridge between our findings and the proposed artefact, and it can be repeated as new scenarios emerge.

4 Results

The following findings in this section are reported from the participants’ perspective on the real-world RE challenges in *HAISS* development. Specifically, how role-specific contexts, recurring failure modes, and different ethical priorities create requirements that cannot be captured by generic, one-size-fits-all ethical frameworks.

Demographics of Participants. To achieve diversity, we recruited 17 participants (8 female, 9 male) across healthcare systems, institutional settings, and cultural backgrounds from China, Australia, and India. The participants worked in IT companies, clinics, hospitals, and universities, and reported 1–24 years of professional experience. Table 1 summarises roles, countries, experience and levels of AI usage; Healthcare professionals (P2–P4, P14–P15), researchers (P1, P5–P8, P12), and software engineers (P9–P11, P13, P16–P17).

4.1 RQ1 (Use cases of AI) Results

We used thematic analysis to identify application areas for each participant role.

4.1.1 Software Engineers’ perspectives

SE-Led Clinical Model Development. Software engineers focused on designing and training domain-specific clinical models rather than deploying generic tools. P9 described augmenting minority-class skin cancer data to address dataset imbalance: *“In the medical domain, we required a balanced dataset, so I used generative AI to enhance the control subject data – we had around 400,000 images used to classify lesions from darker skin as malignant.”* P10 also reported a high staging accuracy of their developed system in cancer detection, though it varied across cancer types.

Leveraging AI Tools as “Hacks” for Quick Explanations. Software engineers used LLMs to bridge knowledge gaps when working across unfamiliar clinical domains. P11 described using chatbots for explanation of multidisciplinary fields: *“When working with medical professionals from a field I am not familiar with, these tools can give you a “hack” or a basic explanation of a topic.”* P17 similarly noted reliance on AI when going through fragmented medical data and ethics approvals.

SE Practices for AI Adoption in Healthcare. Software engineers consistently identified interoperability and data isolation as the primary deployment challenges of *HAISS*. P13 described the burden in the background: *“There is a lot of effort being done in the background to synchronise the data, to fit into every single system, to provide consistency so they get the data rather than manual digging.”* P11 also described that regulatory constraints simultaneously restrict data access and reduce prototype quality: *“Even if you had to do a prototype, it has to be on minimal accuracy and minimal data.”* The conflict between ethical safeguards and system quality is a distinctive challenge for *HAISS*.

4.1.2 Healthcare Professionals’ perspectives

Use of Medical Bots in Practice. Healthcare professionals described AI as embedded across multiple clinical workflow modalities. P2 described two types of bots in Chinese hospitals: surgical support robots for min. invasive procedures and pharmacy dispensing bots that automatically collect and dispense medication based on prescriptions.

Use of AI Assisted Imaging System. P3 described a cloud-based CT diagnostic system: *“Our doctors from the CT department use this system to aid them with the diagnosis of patients’ CT scans. The system detects anomalies and generates a report for the doctors.”* This system draws on large reference databases to support diagnostic uses across orthopedic and radiology departments.

Use of Online Medical Scribe Platforms. P14 described using both Heidi¹ and i-scribe², which are online medical scribe platforms that help transcribe consultations and generate draft clinical letters, reducing administrative burden. P14 emphasised the benefit of these platforms: *“My clinical notes are a lot more accurate and rarely anything is missed”*, while also noting that AI sometimes adds clinical advice that was never given, requiring thorough review of every generated note.

4.1.3 Healthcare Researchers’ perspectives

Use of AI Models for Data Processing. Healthcare researchers predominantly used AI for data-processing tasks, motivated by speed and the need to manage large, heterogeneous datasets. P5 described combining traditional Chinese medicine and Western medicine data using AI models for interpretation, while P6 used large models to process classical medical texts imported from classical sources. These tasks highlight AI as a tool for accelerating data interpretation that researchers had to frequently validate.

Use of AI Models for Image Processing. AI models were also widely used for automated image analysis. P7 described using AI algorithms for human skeletal structure segmentation, automatically generating images in various formats across large visual datasets. P1, specialising in biomechanics and gait motion, similarly used AI models for walking video classification and image processing. For these researchers, AI enabled the processing of large volumes of visual data that would be difficult to handle manually.

¹ <https://www.heidihealth.com/>

² <https://i-scribe.com.au/>

Labelling Datasets Smartly. Healthcare researchers also used AI to label datasets more efficiently, reducing manual annotation effort and improving consistency at scale. P8 stated: *“I use AI models to label my datasets smartly and generate them in various formats”* P12 confirmed good labelling accuracy on omic-based data: *“After I collect and organise my omic-based data, I use AI models to label them – they have a good accuracy in labelling.”*

RQ1 Findings: Software engineers, healthcare professionals and researchers adopt AI for different objectives. Software engineers focus on building customised AI frameworks, solving interoperability problems, and bridging cross-disciplinary knowledge gaps. Healthcare professionals focus on practical clinical applications such as medical bots, AI-assisted imaging, and scribe platforms. Researchers devote most of their effort to data-centric tasks such as augmenting imbalanced datasets, interpreting complex texts, and automating image segmentation. Despite their varied approaches, all three groups share a common goal of improving healthcare outcomes and supporting medical staff.

4.2 RQ2 (Challenges and Limitations) Results

We further asked participants to describe the main challenges and limitations they encounter when applying AI in their clinical, research or *HAISS* development workflows.

Model Reliability & Quality. Healthcare professionals reported diagnostic inaccuracies from AI-assisted imaging systems (P3, P4). An illustrative example came from a clinician using AI scribe tools, where the system added clinical advice that had never been spoken, recommending lifestyle modifications based on the patient’s BMI without the clinician having mentioned it (P14). Software Engineers also highlight hallucinations more broadly: P11 noted that output always requires sanity checking and P9 described persistent garbage outputs even after extensive model training. Across different roles, the unreliability of AI forces practitioner into time-consuming manual verification, reducing the efficiency gains AI was intended to provide.

Data Quality & Processing. Healthcare researchers reported the most acute data quality challenges. P7 described a model that filled gaps in incomplete records with fabricated data, increasing manual review complexity. A particularly striking example came from P9, who trained a skin cancer classifier on 400,000 images drawn predominantly from European and American populations: because the training data contained almost no darker-skinned samples, the model consistently classified benign lesions from Asian skin as malignant – a false positive with significant potential to cause patient distress. P5 reported domain mismatch errors where specialised medical terminology was misrecognised, yielding erroneous outputs. These failures demonstrate how data representativeness and completeness directly affect the reliability and fairness of *HAISS* outputs.

Human Oversight & Skills Gaps. Healthcare professionals consistently described mandatory manual validation as a basic operational necessity. P14 stated clearly: *“I never rely on the answer of AI to a specific scientific question. Rather, I ask AI to provide the evidence and articles. I make the final call.”* The need for medical expert validation also created resource constraints for researchers: P9 noted that generating realistic clinical images still requires domain expert review, which is both costly and time-consuming given the burden already placed on healthcare systems. P16 highlighted that non-technical stakeholders, such

as medical practitioners and researchers, often adopt AI tools without fully understanding their limitations or the associated ethical risks. This places an additional responsibility on SE practitioners to account for gaps in stakeholder awareness during requirements elicitation.

Institutional & Regulatory Constraints. Data access red tape was a consistent barrier described. P13 described a structural paradox in ethics approval processes: “*You can’t even request if you don’t know what you’re requesting for,*” meaning software engineers cannot specify required data fields in their ethics applications because they lack the data visibility to know what to ask for. P9 raised privacy concerns that aligned with reinforced institutional restrictions. P12 observed that commercial AI platforms often operate on tiered pricing models, where paid subscriptions offer stronger data privacy and security features than free access. In healthcare contexts, this means institutions with greater financial resources can provide stronger data protections for their patients, while under-resourced practitioners may inadvertently expose sensitive data to greater risk.

RQ2 Findings: Participant groups reported concerns shaped by their professional roles and goals. Software engineers highlighted technical obstacles, hallucinations, and institutional restrictions that limit AI deployment in healthcare. Healthcare professionals described diagnostic issues, the burden of mandatory manual validation, and AI-generated content that was inconsistent with clinical facts. Healthcare researchers identified fabricated outputs, demographic bias in the training data, and unreliable model behaviour, all of which compromised the validity of the research.

4.3 RQ3 (Ethical Considerations) Results

For RQ3, we asked participants to rank five ethical dimensions on a scale of 1 (highest priority) to 5 (lowest priority), and to explain their reasoning. Privacy received the highest priority (mean = 1.2), followed by safety (mean = 2.2). Bias ranked third (mean = 3.3), while accountability (mean = 3.7) and transparency (mean = 4.5) ranked lowest.

Privacy. emerged as the top priority across all roles. Participants consistently emphasised that patient data must remain tightly controlled, grounding this concern in both institutional policy and personal experience. P14 described stripping patient identifiers from all AI-assisted consultations, while P12 noted that even unpublished research data is strictly prohibited from being entered into AI tools in their institution. In Chinese hospitals, P2 described strict governance measures that prevent the export of patient data outside the hospital.

Safety. ranked second and was also closely linked to privacy from participants’ perspectives. P4 stressed the importance of data containment: “*For imported machines and bots, the database should be based in China*”, which highlights how participants associate safety with physical and institutional control over sensitive data.

Bias. was ranked third. P9 described how datasets trained on different populations can cause patient distress and clinical harm. By contrast, practitioners whose systems undergo institutional validation prior to deployment (P11, P16) considered bias more manageable, since verification processes identify biased outputs before they reach patients.

Accountability. was ranked fourth, and divergent across roles. P17 argued that without clear attribution of responsibility for AI outputs, it is difficult to correct the errors: *“When a ML tool is made, we don’t know who is making that model. If there is anything wrong with a patient’s medication or history, there will be no one accountable.”* P9 similarly argued that given the life-critical stakes, accountability must be the primary concern for developers. Healthcare professionals framed accountability within existing clinical responsibility structures, as doctors are already responsible for decisions made by *HAISS*, making additional accountability feel less urgent. P16 noted that, in their context, accountability was largely pre-empted by design: their application explicitly prepares patients for GP consultations rather than replacing clinical advice, removing ambiguity about who bears responsibility.

Transparency. was ranked the lowest overall. Healthcare professionals stated that clinicians lack the capacity to understand the logic behind *HAISS* outputs. P14 stated that *“can deal with not much transparency, I am 100% sure about the rest.”* P9, a software engineer, stated that while explainability tools such as SHAP and LIME are promising, a significant gap remains in interpreting black-box AI in clinical contexts. P12, declined to provide a single ranking, noting that *“priority depends entirely on context and purpose: for patient care, privacy comes first; for teaching, transparency; for research, security.”* This response illustrates that a single, unified ethical requirements ranking is insufficient, and that role/context-specific elicitation is required.

RQ3 Findings: Privacy and safety were ranked in the top two positions by a large majority of participants and with little spread, indicating that both function as a baseline expectation across all three roles rather than as a discriminating priority; privacy was nonetheless ranked ahead of safety more often than the reverse. Bias, accountability, and transparency were ranked lower on average but with substantially wider variation, and their placement was tied to whether a participant had encountered a concrete failure of that kind. Accountability was ranked lower by healthcare professionals, who described it as already absorbed by existing clinical responsibility structures, than by software engineers, who described it as unassigned. One participant declined to rank the dimensions at all, arguing that priority is determined by purpose rather than by role.

5 Discussion

Our findings suggest that ethical risks in *HAISS* are not uniform. Instead, they vary systematically with (i) practitioners’ roles and responsibilities, (ii) the regulatory constraints they operate under, and (iii) the organisational settings in which *HAISS* are designed and used. This variation explains why “one-size-fits-all” ethical requirements often fail in practice, and why responsible *HAISS* development benefits from cooperative but role-specific RE.

Lesson 1: Ethical requirements in *HAISS* are role-specific, not universal. Across all groups, privacy (mean ≈ 1.2) and safety (mean ≈ 2.2) consistently appear as baseline NFRs; participants widely expect that sensitive data must be protected and that system behaviour should not introduce avoidable harm. The remaining dimensions, however, showed greater variance across roles, and their importance depended on how each role encountered ethical concerns in everyday work. Healthcare professionals concentrated on issues that directly affect clinical decision-making, researchers emphasised reproducibility, and software engineers

■ **Table 2** Role-specific If-Then patterns derived from practitioner perspectives.

If (Scenario)	Then (Response)	Role	Guideline(s)
Building models for underrepresented populations with imbalanced data (P9).	Augment minority classes, document in model cards, validate with domain experts, and report metrics separately for synthetic and natural data.	SE	EU AI Act [11]; FDA GMLP [31]
SE teams lack medical knowledge and use LLMs for quick explanations (P11, P17).	Limit LLM use to preliminary understanding, maintain validated glossaries, and mandate clinical review before implementation.	SE	IEEE EAD [6]; ACM Ethics [1]
Deploying AI where data must stay within institutional boundaries (P13, P9).	Architect for on-premise or private cloud only, block public APIs, and log all access to verify no external transmission.	SE	HIPAA [28]; GDPR Art. 44–50 [10]; PIPL [26]
AI imaging systems produce occasional diagnostic errors (P3, P4).	Require radiologist double-read of AI-flagged findings, surface confidence scores, and maintain an audit trail.	HP	ACR [3]; FDA CDS [30]
Clinical staff lack the knowledge to operate AI tools (P2).	Provide mandatory role-specific training with certification and simplified interfaces, and collect usability feedback.	HP	WHO [33]; FDA HF [29]
AI fabricates data to fill gaps (P7).	Reject incomplete records, use statistical rather than generative imputation, mark synthetic values, and exclude them from primary analysis.	HR	FAIR [32]; CONSORT-AI [21]

SE = Software Engineer; HP = Healthcare Professional; HR = Healthcare Researcher.

focused on implementable governance mechanisms and on how ethical requirements translate into testable constraints. For RE, this implies that eliciting ethics requirements in *HAISS* cannot rely on a single fixed checklist, but must be stratified by role. We state this as a hypothesis rather than a finding. Because role and country are confounded in our sample, the stratification we observe may be organised as much by regulatory setting as by professional role. What the data more securely support is the weaker claim that a single, fixed checklist did not match how any of our participant groups encountered ethical concerns.

Lesson 2: Role boundaries are the primary hotspots for ethical risk. Several severe issues reported were not confined to a single role; they appeared at the boundaries, where one group produces artefacts that another group relies on. Examples included hallucinated outputs that appear clinically plausible and are mistakenly trusted; demographic bias that increases false positives and forces clinicians to make rapid judgments under pressure; and fabricated or low-quality data propagating through research workflows. These failures often arise because validation expectations and accountability are not explicitly specified at handoffs, allowing errors or uncertainty to flow downstream. This suggests that handoffs themselves should be treated as requirements artefacts, with explicit validation criteria and accountability assigned at each role boundary.

Lesson 3: Regulatory compliance and ethical requirements are related, but not interchangeable. Compliance is essential, but it represents a minimum legal threshold rather than a complete account of ethical practice. A system may satisfy data residency rules while

still producing biased outcomes, or obtain consent in a legally correct way while operating as a “black box” that makes accountability difficult to assign when something goes wrong. For SE practitioners, this is why data governance and residency should be treated as core architectural requirements rather than background tasks, with provenance tracking and automated completeness checks built in, rather than relying on compliance sign-off alone.

Based on our findings, we propose role-specific If-Then patterns as lightweight SE artefacts that operationalise practitioner-identified ethical priorities into actionable requirements. Each pattern specifies a trigger condition (If) drawn from a recurring scenario in our interview data, a concrete SE response (Then) grounded in both participant suggestions and established regulatory guidance, the responsible role, and the applicable guideline. The patterns were derived by mapping each RQ2 challenge sub-theme to the ethical dimension it most directly threatens (RQ3), then formulating a response that combines participant suggestions with the corresponding regulatory obligation. Table 2 presents an initial set of six patterns spanning all three roles, intended as structured prompts during requirements elicitation. We did not attempt to operationalise every RQ2 challenge. A sub-theme was carried forward only if it was reported by at least two participants or by one participant describing a specific incident, if its trigger could be stated as a condition a development or clinical team can observe, and if a response could be formulated that maps onto an obligation in a publicly available instrument. Sub-themes excluded on these grounds were P13’s ethics-approval paradox, whose trigger is organisational rather than a property of the software, and P5’s terminology mismatch, for which we could not formulate a response that did not restate the problem. Guidelines were selected by matching each response to the narrowest instrument that covers the obligation and applies in at least one of the three jurisdictions in our sample; we do not claim that the cited instrument mandates the specific response, only that the response is intended to discharge the obligation it states. The responses combine participant suggestions with our own additions, most often a verification or logging step that makes a participant’s suggestion checkable, so the patterns are our formulation rather than a direct transcription of participant views. Their triggers are grounded in observed practice, but the responses have not been applied to a real project, and evaluating them through practitioner workshops is the first item of our future work.

6 Threats to Validity

6.1 Construct Validity

The five ethical dimensions were not derived inductively from the interview data. We adopted them from our prior systematic mapping study of ethical concerns in generative AI [14], which synthesised 39 primary studies into five recurring dimensions. Because the same team produced the framework and applied it as an elicitation instrument, a concern absent from the mapping study is unlikely to have surfaced here. The ranking task compounds this. It captures relative priority only, so a low rank means a participant placed that dimension below the others, not that they judged it unimportant, and the low rank of transparency should be read in this way. It also records what participants say they prioritise rather than what their projects protect in practice, since we did not observe any participant’s work. To limit these effects, each interview included open discussion in which participants could raise concerns outside the five labels, and we asked every participant to justify their ordering, which is how the qualitative account in Section 4.3 was built. In the planned practitioner workshops we will elicit priorities against concrete scenarios rather than in the abstract, and compare stated priorities against project artefacts such as requirements documents and ethics applications.

6.2 Internal Validity and Credibility

The primary coding was performed by a single researcher, which is the most significant threat to the credibility of our themes. Section 3 describes the steps we took to reduce it: an audit trail linking every code to the transcript segment it came from, and explicit attention to cases that diverged from the emerging pattern. We did not conduct a formal inter-coder reliability assessment, nor did we ask participants to validate the themes. Our themes should therefore be read as an interpretation that can be traced back to the data and justified, rather than as the only structure a coder could have derived from these transcripts. To make competing interpretations possible, we release the codebook with links from each code to its source segment, so that another researcher can trace how each theme was built and construct an alternative reading from the same material.

6.3 External validity and Transferability

Our sample of 17 participants is small and unevenly distributed, and the distribution is not independent of role. Healthcare professionals were recruited only in China and Australia; software engineers, only in India and Australia; and India is represented by a single participant. Role and country are therefore confounded: any pattern we attribute to the software engineering role is, on this sample, indistinguishable from one attributable to the regulatory or institutional context in which those engineers work. This affects RQ3 most directly, since data-protection regimes differ sharply across the three jurisdictions. Australia is the only country represented in all three role groups, so we checked whether the role-level ordering holds there as well. Privacy and safety retain their positions, but the placement of bias, accountability, and transparency differs from the overall pattern, so the role differences we report for these three dimensions may partly reflect jurisdictional differences. Recruitment through professional and social platforms further biases the sample toward practitioners already engaged with AI. We therefore make no claim of statistical generalisation and present the role-specific findings as hypotheses. To test them we are extending the study to a sample stratified by role within each jurisdiction, so that role and country can be varied independently, and we will re-examine the ordering reported here before generalising from it.

7 Conclusion

In this paper, we demonstrate that understanding multifaceted AI adoption in healthcare requires considering the perspectives of stakeholder groups with distinct roles. By examining RQ1, we found that healthcare researchers use AI for data-centric tasks, healthcare professionals focus on integrating AI into diagnostic workflows, and software engineers focus on building and maintaining AI systems. Each group has unique challenges they encounter (RQ2) and ethical considerations they prioritise (RQ3), which are shaped by their objectives.

Based on these findings, we suggest creating role-specific guidelines that ground high-level principles in everyday practices by formulating if-then patterns. While our sample captures diverse geographic and institutional contexts, it remained limited in size and scope. As future work, we plan to implement these if-then patterns within role-specific guidelines, drawing on existing regulatory frameworks and ethical standards, and to validate and refine them across a wider range of healthcare environments to ensure broader applicability and relevance. By incorporating continuous practitioner feedback, we aim to produce a set of actionable rules that can be streamlined and, ultimately, embedded into an adaptive, ethical healthcare AI framework capable of evolving with emerging technologies and practical scenarios.

Data Availability Statement

An anonymised replication package containing the interview protocol and the thematic code-book is available at https://anonymous.4open.science/r/EMSE_Emerging-Results-1F8B

References

- 1 Association for Computing Machinery. ACM Code of Ethics and Professional Conduct, June 2018. Adopted by ACM Council June 22, 2018. URL: <https://www.acm.org/code-of-ethics>.
- 2 Walaa Badawy and Mostafa Shaban. Exploring geriatric nurses' perspectives on the adoption of ai in elderly care a qualitative study. *Geriatric Nursing*, 2025.
- 3 Adrian P Brady, Bibb Allen, Jaron Chong, Elmar Kotter, Nina Kottler, John Mongan, Lauren Oakden-Rayner, Daniel Pinto Dos Santos, An Tang, Christoph Wald, et al. Developing, purchasing, implementing and monitoring ai tools in radiology: practical considerations. a multi-society statement from the acr, car, esr, ranzer & rsna. *Canadian Association of Radiologists Journal*, 2024.
- 4 Lukas Buess, Matthias Keicher, Nassir Navab, Andreas Maier, and Soroosh Tayebi Arasteh. From large language models to multimodal ai: A scoping review on the potential of generative ai in medicine. *Biomedical Engineering Letters*, 2025.
- 5 Danton S Char, Nigam H Shah, and David Magnus. Implementing machine learning in health care—addressing ethical challenges. *New England Journal of Medicine*, 2018.
- 6 Raja Chatila and John C Havens. The ieee global initiative on ethics of autonomous and intelligent systems. In *Robotics and well-being*. Springer, 2019.
- 7 Carter Cousineau, Rozita Dara, and Ataharul Chowdhury. Trustworthy ai: Ai developers' lens to implementation challenges and opportunities. *Data and Information Management*, 2025.
- 8 Daniela S Cruzes and Tore Dyba. Recommended steps for thematic synthesis in software engineering. In *2011 international symposium on empirical software engineering and measurement*. IEEE, 2011.
- 9 Cansu Yüksel Elgin and Ceyhun Elgin. Ethical implications of ai-driven clinical decision support systems on healthcare resource allocation: a qualitative study of healthcare professionals' perspectives. *BMC Medical Ethics*, 2024.
- 10 European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation), 2016.
- 11 European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), July 2024.
- 12 Magali Goirand, Elizabeth Austin, and Robyn Clay-Williams. Implementing ethics in healthcare ai-based applications: a scoping review. *Science and Engineering Ethics*, 2021.
- 13 Dennis Henzler, Sebastian Schmidt, Ayca Koçar, Sophie Herdegen, Georg L Lindinger, Menno T Maris, Marieke AR Bak, Dick L Willems, Hanno L Tan, Michael Lauerer, et al. Healthcare professionals' perspectives on artificial intelligence in patient care: a systematic review of hindering and facilitating factors on different levels. *BMC Health Services Research*, 2025.
- 14 Yutan Huang, Chetan Arora, Wen Cheng Huong, Tanjila Kanij, Anuradha Madugalla, and John Grundy. Ethical concerns of generative ai and mitigation strategies: A systematic mapping study. *Applied Soft Computing*, 2026.
- 15 Sandra LJ Johnson. Ai, machine learning, and ethics in health care. *Journal of Legal Medicine*, 2019.
- 16 Golnar Karimian, Elena Petelos, and Silvia MAA Evers. The ethical issues of the application of artificial intelligence in healthcare: a systematic scoping review. *AI and Ethics*, 2022.
- 17 Mohamed Khalifa and Mona Albadawy. Ai in diagnostic imaging: revolutionising accuracy and efficiency. *Computer Methods and programs in biomedicine update*, 2024.

- 18 Sunita Krishnan, Michael Matheny, Elaine Fontaine, and Laura Adams. An artificial intelligence code of conduct for health and medicine: Essential guidance for aligned action, 2025.
- 19 Tiffany I Leung, Andrew J Coristine, and Arriel Benis. Ai scribes in health care: Balancing transformative potential with responsible integration. *JMIR Medical Informatics*, 2025.
- 20 Fan Li, Nick Ruijs, and Yuan Lu. Ethics & ai: A systematic review on ethical concerns and related strategies for designing with ai in healthcare. *Ai*, 2022.
- 21 Xiaoxuan Liu, Samantha Cruz Rivera, David Moher, Melanie J Calvert, Alastair K Denniston, Hutan Ashrafian, Andrew L Beam, An-Wen Chan, Gary S Collins, Ara DarziJonathan J Deeks, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the consort-ai extension. *The Lancet Digital Health*, 2020.
- 22 Lena Pons and Ipek Ozkaya. Priority quality attributes for engineering ai-enabled systems. *arXiv preprint arXiv:1911.02912*, 2019. [arXiv:1911.02912](https://arxiv.org/abs/1911.02912).
- 23 Osama Mohamed Elsayed Ramadan, Majed Mowanes Alruwaili, Abeer Nuwayfi Alruwaili, Mohamed Gamal Elsehrawy, and Sulaiman Alanazi. Facilitators and barriers to ai adoption in nursing practice: a qualitative study of registered nurses’ perspectives. *BMC nursing*, 2024.
- 24 Rakesh Kumar Sahoo, Sapna Negi, Santos Kumar Baliarsingh, Bhuputra Panda, Sanghamitra Pati, et al. Health professionals’ perspectives on the use of artificial intelligence in healthcare: A systematic review. *Patient Education and Counseling*, 2025.
- 25 Melanie Smallman. Multi scale ethics—why we need to consider the ethics of ai in healthcare at different scales. *Science and Engineering Ethics*, 2022.
- 26 Standing Committee of the National People’s Congress. Personal Information Protection Law of the People’s Republic of China. Presidential Order No. 91, August 2021. Adopted August 20, 2021; Effective November 1, 2021. URL: http://en.npc.gov.cn.cdurl.cn/2021-12/29/c_694559.htm.
- 27 Eric Topol. *Deep medicine: how artificial intelligence can make healthcare human again*. Hachette UK, 2019.
- 28 U.S. Department of Health and Human Services. Standards for privacy of individually identifiable health information (HIPAA Privacy Rule). 45 C.F.R. §§ 160, 164, 2002. URL: <https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html>.
- 29 U.S. Food and Drug Administration. Applying Human Factors and Usability Engineering to Medical Devices: Guidance for Industry and Food and Drug Administration Staff. Guidance document, U.S. Food and Drug Administration, Center for Devices and Radiological Health, February 2016. Document Number 1757; Docket FDA-2011-D-0469. URL: <https://www.fda.gov/media/80481/download>.
- 30 U.S. Food and Drug Administration. Clinical Decision Support Software: Guidance for Industry and Food and Drug Administration Staff. Technical Report Document No. 1400062, U.S. Food and Drug Administration, September 2022. Docket FDA-2017-D-6569. URL: <https://www.fda.gov/media/109618/download>.
- 31 U.S. Food and Drug Administration, Health Canada, and Medicines and Healthcare products Regulatory Agency. Good Machine Learning Practice for Medical Device Development: Guiding Principles. Technical report, U.S. Food and Drug Administration, October 2021. Joint guidance from FDA, Health Canada, and UK MHRA. URL: <https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles>.
- 32 Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan Willem Boiten, Luiz Bonino Da Silva Santos, Philip E Bourne, et al. The fair guiding principles for scientific data management and stewardship: Comment. *Scientific data*, 2016.
- 33 World Health Organization. WHO guideline: Recommendations on digital interventions for health system strengthening: evidence and recommendations. Guideline WHO/RHR/19.10, World Health Organization, Geneva, 2019. Available under CC BY-NC-SA 3.0 IGO. URL: <https://iris.who.int/items/cc58cecd-77ac-412a-a9fd-4752b11e5d91>.