

Stakeholder-centered design of explainable AI for MRI analysis in multiple sclerosis: Generative insights toward clinical integration

Margherita Motta^{*1}, Wen Zhan¹, Nataliia Molchanova^{2,3,4,5}, Federico Spagnolo^{6,7,8}, Sebastian Baez-Lugo¹, Clara Evans¹, Pedro Macias Gordaliza^{5,3,2}, Alessandro Cago^{6,9,8,10}, Vincent Dunet¹¹, Silvia Pistocchi¹¹, Matthias Anthony Mutke¹², Lluís Borràs Ferris^{4,13}, Beatrice Gobbo¹⁴, Mauricio Reyes^{15,16}, Esther Ruberte^{6,8}, Frederic Erard^{17,18}, Nicolas Henchoz¹, Henning Müller^{4,19,20}, Cristina Granziera^{6,9,8,10}, Adrien Depeursinge^{4,13}, Merixtell Bach Cuadra^{5,3,2}, Delphine Ribes¹

¹ EPFL+ECAL Lab, École polytechnique fédérale de Lausanne (EPFL), Lausanne, Switzerland

² Faculty of Biology and Medicine, University of Lausanne (UNIL), Lausanne, Switzerland

³ Radiology Department, Lausanne University Hospital (CHUV), Lausanne, Switzerland

⁴ Institute of Informatics, University of Applied Sciences Western Switzerland (HES-SO Valais), Sierre, Switzerland

⁵ CIBM Center for Biomedical Imaging, Lausanne, Switzerland

⁶ Translational Imaging in Neurology (ThINK) Basel, Department of Medicine and Biomedical Engineering, University Hospital Basel and University of Basel, Basel, Switzerland

⁷ Department of Neurology, University Hospital Basel, Basel, Switzerland

⁸ Research Center for Clinical Neuroimmunology and Neuroscience Basel (RC2NB), University Hospital Basel and University of Basel, Basel, Switzerland

⁹ Multiple Sclerosis Centre, Departments of Neurology, Clinical Research and Biomedicine, University Hospital and University Basel, Switzerland

¹⁰ Department of Health Sciences, University of Genova, Genova, Italy

¹¹ Department of Medical Radiology, Service of Diagnostic and Interventional Radiology, Neuroradiology Unit, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland

¹² Department of Radiology and Nuclear Medicine, Division of Neuroradiology, University Hospital Basel, Basel, Switzerland

¹³ Nuclear Medicine and Molecular Imaging Department, Lausanne University Hospital (CHUV), Lausanne, Switzerland

¹⁴ Politecnico di Milano, Department of Design, Milan, Italy

¹⁵ ARTORG Center for Biomedical Engineering Research, University of Bern, Bern, Switzerland

¹⁶ Department of Radiation Oncology, Inselspital, Bern University Hospital and University of Bern, Bern, Switzerland

¹⁷ Faculté de droit, des sciences criminelles et d'administration publique, Lab' Santé et droit, Université de Lausanne, Lausanne, Switzerland

¹⁸ Faculté de biologie et de médecine, Institut des Humanités en Médecine, Université de Lausanne, Lausanne, Switzerland

¹⁹ The Sense Innovation and Research Center, Sion, and Lausanne, Switzerland

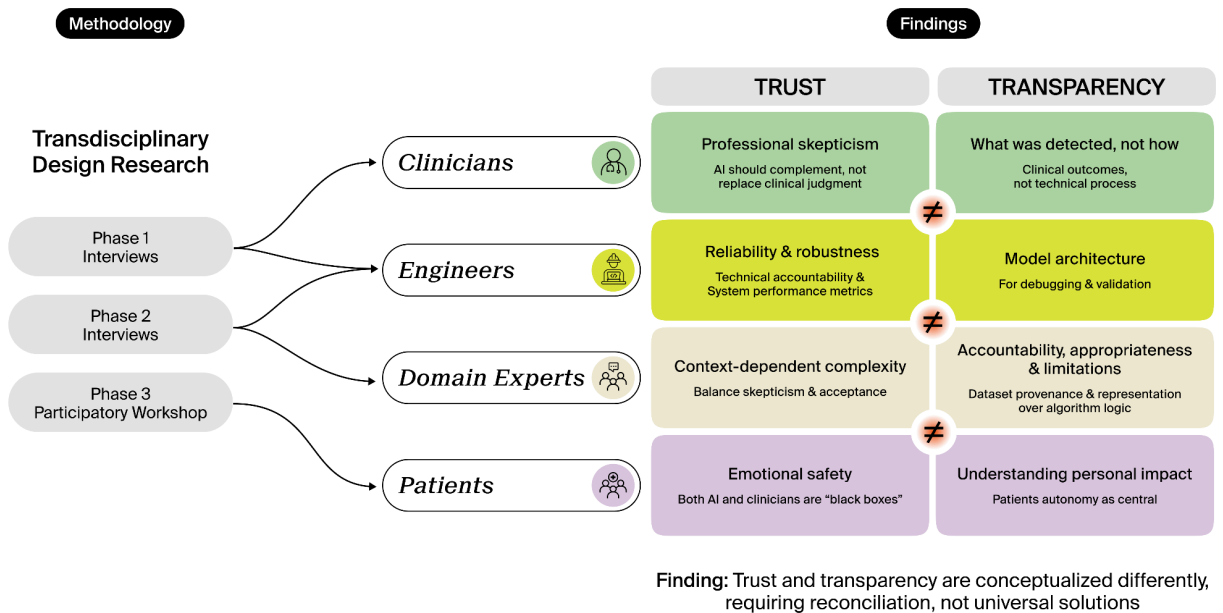
²⁰ Medical faculty, University of Geneva, Geneva, Switzerland

* Corresponding author
E-mail: margherita.motta@epfl.ch

Abstract

Explainable AI (XAI) techniques have emerged to address the black-box nature of deep learning models in medical imaging, yet current research focuses primarily on algorithmic development rather than stakeholder engagement. This study investigates how diverse stakeholders conceptualize trust, transparency, and safety in XAI systems for Multiple Sclerosis (MS) radiology through transdisciplinary design research. We conducted a three-phase generative study engaging radiologists and neurologists (N=6), technical experts (N=5), domain experts from design, ethics, law, and AI research (N=6), and lay public participants (N=6). Methods included semi-structured interviews and a participatory workshop employing creative materials to surface embodied perspectives about vulnerability and AI in healthcare.

Our findings reveal fundamental misalignments between technical XAI assumptions and stakeholder realities. Trust is conceptualized fundamentally differently across professional communities, requiring reconciliation rather than calibration: clinicians frame trust as skepticism and verification, engineers as technical accountability, and domain experts as context-dependent complexity. Patients experience a "double black box" where both AI systems and clinicians operate opaquely, with emotional safety emerging as integral to patient safety rather than secondary to technical error prevention. Context-dependence operates as a multilevel structural constraint (clinical, regulatory, and technical) that shapes adoption possibilities regardless of algorithmic sophistication. We derive design implications for heterogeneous trust frameworks, context-adaptive interfaces, and patient-facing transparency, demonstrating design research's capacity to surface stakeholder knowledge invisible to technical-only approaches.



Author Summary

Explainable AI (XAI) techniques are being developed to make deep learning models more interpretable for clinical use, yet most research focuses on algorithmic performance rather than engaging the diverse stakeholders who will interact with these systems. We conducted a three-phase study within a transdisciplinary consortium developing AI for Multiple Sclerosis brain imaging, combining interviews with clinicians, engineers, and domain experts, alongside a participatory workshop with lay participants exploring vulnerability in AI-assisted healthcare. Our findings reveal that trust is not a single construct to calibrate but is conceptualized in fundamentally different ways across professional communities: as verification practice by clinicians, technical accountability by engineers, and context-dependent complexity by domain experts. Workshop participants identified what we term a "double black box": patients already experience clinical reasoning as opaque, and AI introduces a second layer of opacity rather than

originating the problem. Furthermore, context-dependence operates as a multilevel structural constraint like clinical workflows, regulatory requirements, and technical infrastructure each shape adoption possibilities independently of algorithmic sophistication. These results demonstrate that meaningful clinical integration of XAI requires engaging the perspectives of all those affected, and that design research methods can surface knowledge invisible to purely computational approaches.

1 Introduction

Deep learning has transformed medical imaging, improving automated detection and lesion delimitation (segmentation) tasks that previously required extensive manual analysis (1). However, the black-box nature of these models creates interpretability challenges (2), particularly in high-stakes clinical domains where decision making directly affects patient outcomes (3). Explainable AI (XAI) techniques including saliency maps, attention mechanisms, and uncertainty quantification - have emerged to address this opacity by providing insights into model reasoning (4,5).

Multiple Sclerosis (MS) presents a particularly critical case for examining AI integration in medical imaging. As a chronic neurological condition affecting approximately 2.8 million people globally, MS requires ongoing monitoring through repeated MRI scans to assess disease progression and treatment efficacy (6). For patients - predominantly diagnosed between ages 20-50 - this chronic condition brings significant psychological burden, with studies reporting elevated anxiety

(34-40%) and distress, particularly around diagnostic uncertainty and unpredictable disease progression (7,8). The McDonald criteria, used internationally for MS diagnosis, integrates MRI evidence - alongside clinical and laboratory findings - to demonstrate lesion dissemination across different brain regions and at different timepoints (9). However, detailed MRI lesion analysis in MS remains time-intensive and subject to significant inter-rater variability, as radiologists must manually examine a large number of slices to segment and measure individual lesions (10). This combination of time-consuming pattern recognition, high imaging volumes, and known variability makes MS imaging well-suited to AI assistance, yet the stakes remain high. Diagnostic decisions shape treatment plans, while disease monitoring informs disease-modifying treatment adjustments that carry significant side effects and long-term health implications. The chronic nature of MS requires patients to maintain trust in the diagnostic and monitoring process across years of repeated imaging and clinical encounters (11,12).

AI-assisted analysis of MS MRI scans serves as a tool to ensure consistent and semi-automated assessment, with XAI offering interpretable solutions that clarify how models derive their outputs. Recent advances include uncertainty quantification - methods that distinguish between what AI predicts (e.g., lesion detection) and how confident the model is in that prediction - helping clinicians assess prediction reliability (13). However, current XAI research remains largely focused on algorithmic advancements evaluated through computational metrics, rather than on engagement with end-users in real-world clinical workflows (14,15). Even studies involving clinicians rarely expand to other stakeholders, leaving out

patients (16) as well as ethicists, legal scholars, and designers, largely overlooking their perspectives on accountability, usability, and wellbeing.

Our work directly addresses this gap through the MSXplain transdisciplinary research consortium, developing XAI systems for MS analysis. As design researchers collaborating with technical developers and clinical partners, we employed qualitative and participatory methods to examine different perspectives that inform XAI development, investigating not merely technical performance and algorithmic accuracy, but how diverse stakeholders conceptualize trust, transparency, and wellbeing when AI enters clinical practice. Design research's methodological pluralism - accessing different forms of knowledge through multiple inquiry methods - proved essential for navigating the distinct mental models, value systems, and priorities operating across disciplines (17,18). This paper reports findings from a three-phase generative study investigating stakeholder perspectives on trust, transparency, and clinical needs: (1) semi-structured interviews with radiologists/neurologists and technical experts (N=11); (2) semi-structured interviews with domain experts in design, ethics, law, AI research, and medical imaging annotation (N=6); and (3) a participatory workshop exploring bodily vulnerability and AI in healthcare with lay public participants (N=6). Preliminary findings from the participatory workshop were previously reported in (19). This paper extends that work through comprehensive three-phase analysis integrating clinical, technical, and experiential perspectives.

This paper makes three contributions to understanding XAI implementation in MS clinical/research contexts. First, we reveal fundamental conceptual misalignments between technical XAI assumptions and stakeholder realities: trust is

conceptualized heterogeneously across professional communities, requiring reconciliation, not calibration. Second, we identify the "double black box" phenomenon where patients experience opacity from both AI systems and clinicians, challenging foundational assumptions that transparency problems originate solely with AI and repositioning explainability as an infrastructure challenge spanning entire care pathways. Third, we demonstrate how context-dependence operates as a multilevel structural constraint - clinical, regulatory, and technical - that shapes adoption possibilities regardless of algorithmic sophistication, exposing tensions between verification requirements and efficiency narratives. From these findings, we derive design implications requiring context-adaptive interfaces that support heterogeneous trust frameworks simultaneously, differentiated stakeholder transparency, and patient-facing communication strategies extending beyond clinician-algorithm interactions.

2 Related work

2.1 Trust, Transparency, and Appropriate Reliance in Clinical AI

Trust in AI-assisted diagnosis is conceptualized as a multifaceted psychological mechanism that mediates clinicians' willingness to rely on AI systems under uncertainty (20–22). In clinical AI contexts specifically, researchers distinguish between trust (an attitude held by the user) and trustworthiness (a property inherent to the AI system) (23). A system can be trustworthy without gaining user trust, or a user might trust a non-trustworthy system due to factors like interface quality, leading to misuse or disuse (24). For effective human-AI collaboration, the goal is not merely to maximize trust, but to achieve calibrated trust or appropriate

reliance (21,25). In fact, maximizing trust can lead to harmful consequences, especially in life-critical applications, if the user accepts all AI recommendations without scrutiny, in process defined automation bias (26). Appropriate reliance requires the human user to apply skepticism and rely on the AI only when its capability justifies it. This calibration is particularly important on a case-by-case, prediction-specific level, helping users correctly accept correct suggestions and reject errors (27). A major barrier to clinical adoption is the inherent lack of transparency and interpretability associated with complex AI models, often referred to as "black-box", where the reasoning behind their predictions remains opaque (28,29). Since medical professionals rely on transparent, evidence-based reasoning, they are often reluctant to rely on systems they cannot fully understand or scrutinize (30).

Explainable AI has emerged as the principal solution to bridge the gap between model performance and human interpretability (28). XAI encompasses methodologies aimed at enabling end-users to comprehend and interpret the outputs and predictions of AI models (14,31,32). Regulatory frameworks reinforce this push toward explainability: the European Union's General Data Protection Regulation (GDPR) emphasizes the need for explainability in automated decision-making (33,34), while the EU Artificial Intelligence Act establishes requirements for transparency, explainability, and human oversight specifically for high-risk AI systems including medical devices (35). Human-centered evaluation of XAI systems with clinical users is an emerging field, with growing evidence on the importance of usability, stakeholder involvement, and context-specific design (34,36,37).

However, current XAI approaches more typically prioritize technical interpretability - targeting developers and engineers seeking to optimize model performance and efficiency - over clinical utility, creating a fundamental disconnect between what systems can explain and what clinicians need to know (38). Existing explainability methods have been found to be insightful often only to algorithmic experts, failing to address clinicians' actual needs for actionable, workflow-aligned explanations (39). Yang et al. demonstrated that AI explanations commonly fail to provide actionable information, with the lack of a shared source of truth between AI and clinician transforming the process into opaque "AI explanation sense-making" rather than clinical decision-making (27). This disconnect extends to workflow complexity: XAI research remains focused on single-modality, single-timepoint explanations despite clinical decision-making requiring multimodal data integration across time (40).

Temporal and organizational constraints further limit XAI's clinical utility. Clinicians' limited time makes in-the-moment trust determinations nearly impossible, with evidence suggesting they prefer to determine trust in a technology once - through evidence-based validation - rather than at each decision point (41). Fogliato et al.'s study with 19 veterinary radiologists demonstrated that workflow design significantly affects clinicians' reliance on AI, particularly under time pressure (42).

In this landscape, research consensus points toward abandoning universal XAI solutions in favor of context-dependent transparency calibrated to specific clinical situations, stakeholder needs, and workflow constraints (43–45).

2.2 Stakeholder Engagement and Participatory Practices in AI Design

The field of AI has undergone a "participatory turn" in recent years, driven by growing recognition that those affected by AI systems should participate in their design (46). This shift responds to documented harms emerging across security, justice, employment, and healthcare domains, revealing that technically-focused, non-participatory development practices cannot meet the optimistic vision of AI intended to enhance human agency and wellbeing (47). Participatory approaches promise multiple benefits: incorporating diverse perspectives that technical experts might miss, representing needs of historically marginalized communities, and enabling AI systems to better reflect stakeholder values and preferences (48).

Healthcare contexts have received particular attention, with researchers investigating stakeholder-driven approaches to explainable AI that prioritize human-AI team performance over purely technical metrics (49). Medical imaging and radiotherapy fields specifically emphasize the necessity of multidisciplinary teams and multiagency approaches for successful AI implementation, recognizing how cross-professional collaboration is essential for addressing implementation barriers (50). This interdisciplinary imperative extends to establishing shared vocabularies; Graziani et al. propose a global taxonomy unifying interpretability terminology across technical and social sciences, recognizing that effective collaboration in high-stakes domains like healthcare requires practitioners, ethicists, and technologists to speak a common language about AI transparency and explainability (51). Gonzalez-Gonzalo et al. (24) further document how numerous AI systems proposed for ophthalmology struggle with real-world implementation despite technical sophistication, highlighting that trustworthiness,

stakeholder engagement, and integration challenges must be addressed alongside algorithmic performance.

However, for researchers and practitioners interested in participatory methods, it remains challenging to assess whether any given approach grants substantive agency to stakeholders or stay merely performative (52,53). This ambiguity matters because participatory claims often mask underlying power asymmetries: development processes systematically privilege technical and institutional voices while marginalizing those most impacted by algorithmic systems (48,54).

Transdisciplinary collaboration - essential for meaningful participation - faces practical obstacles including divergent research paradigms, conflicting incentive structures, and resource constraints (55).

Critical perspectives question whether participation genuinely contributes to responsible AI or legitimizes predetermined trajectories. Lysen and Wyatt (56) provocatively ask whether patient engagement in healthcare AI truly empowers or merely extracts knowledge while preserving existing power relations, cautioning against "refusing participation" that becomes coercive. This critique highlights persistent ambiguity around what constitutes meaningful participation versus symbolic, superficial involvement.

Despite growing consensus on participation's importance, clarity remains elusive regarding operationalization: who should participate, at which development stages, through what mechanisms, and with what authority (47,57).

3 Methods

This study employs transdisciplinary generative research to investigate stakeholder perspectives on explainable AI in MS imaging during early-stage technology development. Our approach recognizes that understanding complex socio-technical systems - particularly in the context of clinical AI implementation - requires engaging diverse stakeholder perspectives, each bringing situated forms of expertise, values, and experiences (58).

Ethical approval for this study was granted by the EPFL Human Research Ethics Committee. All participants provided informed consent after receiving detailed information about the study, data usage and anonymization.

3.1 Multi-Method Design: Overview

Our goal in this exploratory phase was to investigate adoption barriers and requirements for integrating deep learning-based models into MS clinical workflows, with a particular focus on how XAI could mitigate those barriers.

We conducted a three-phase generative study over 12 months, selecting methods that would reveal different dimensions of relationship with AI, trust and transparency, and across diverse stakeholder groups. Our sampling strategy employed purposive sampling throughout all phases, prioritizing depth of insight and stakeholder diversity over statistical representativeness - an approach consistent with exploratory qualitative inquiry aimed at theory generation rather than hypothesis testing (59). Phase 1 engaged clinical and technical experts

through semi-structured interviews (11 participants). Phase 2 expanded the stakeholder ecosystem by interviewing domain experts from design, ethics, law, and AI research (6 participants). Phase 3 employed participatory workshop methods with lay public participants (6 participants), creating space for embodied and metaphorical exploration of vulnerability in AI-assisted healthcare.

With this multi-method approach, our aim was to generate rich, actionable insights that could inform future XAI design and implementation in ways that technical evaluation alone cannot capture.

Phase	Method	N Participants	Data Type	Timeframe
Phase 1	Semi-structured Interviews	11 (E1-E5: engineers, C1-C6: clinicians)	Audio, transcripts, coding	June-September 2023
Phase 2	Semi-structured Interviews	6 (S1-S6: diverse experts)	Audio, transcripts, coding	October - December 2024
Phase 3	Participatory Workshop	6 (WS1-WS6: designers/engineers)	Audio, notes, photos, artifacts	December 2024
TOTAL		23 participants	Mixed Methods	

Table 1 - Summary of Research Phases, Participant Composition, and Data Collection Procedures

3.2 Phase 1: Clinical and Technical Perspectives

The first phase sought to understand how those most directly involved in developing and deploying AI for MS imaging conceptualize trust, transparency, and utility. We recruited 6 clinicians (including neuroradiologists and general radiologists working with MS cases, representing a mix of junior and senior experience levels) and 5 technical experts (4 from research institutions and 1 from industry, all with expertise in AI/ML for medical imaging, summarized in Table 2). Recruitment occurred through our project consortium partners and professional networks, using purposive sampling to ensure diversity in experience levels and subspecialties. Interviews were conducted in English and French based on participant preference. All took place remotely via Zoom, recorded with consent, and lasted 45-60 minutes.

We developed separate semi-structured interview protocols for clinicians and engineers, tailored to their respective expertise while maintaining common themes across both groups. The Phase 1 protocol was intentionally open-ended to allow exploratory conversation. Core themes addressed across both protocols included current MS imaging workflows and tool usage, attitudes toward AI and deep learning in healthcare contexts, factors influencing trust, transparency needs and anticipated challenges and opportunities for automation. Radiologist-specific questions explored clinical workflow, diagnostic challenges, and tool evaluation, with questions like "What tools do you use for diagnosis and monitoring? What do you think of them?" Engineer-specific questions focused on tool development, clinical integration, and physician perspectives, with questions like "How do you

think deep-learning tools can be integrated in the clinical workflow? What are the challenges in developing such tools?"

This semi-structured protocol allowed participants to elaborate on details that came to mind naturally, enabling us to identify both convergent concerns and divergent priorities.

Code	Role	Specialization	Key details
E1	PhD Candidate/Engineer	Uncertainty quantification, MS lesion segmentation	Focus: model uncertainty, metrics, clinical workflow understanding
E2	PhD Candidate/Engineer	Medical image analysis	Focus: lesion detection, segmentation, research vs clinical tasks
E3	Engineer/Researcher	AI/ML methods	Focus: algorithm testing, population data challenges
E4	Engineer/Researcher	Clinical integration, e-health	Focus: bridging algorithms to usable interfaces, report generation
E5	Team Lead Engineer	Pre-development R&D, MRI	Senior engineer, industry perspective, workflow integration

		sequences & post-processing	
C1	Neurologist/MD	MS, brain atrophy, neuro-oncology	MS research focus
C2	Radiologist/MD	MS Imaging	Experienced radiologist
C3	Radiologist/MD	MS imaging, neuroradiology	Focus: reading time variation, trust building
C4	Radiologist/MD	MS imaging, neuroradiology	Focus: verification workflow, human oversight necessity
C5	Chief Neuroradiologist	Neuroradiology, neurodegenerative diseases, MS	Senior clinician, discusses confidence scores
C6	Neurologist/Radiologist	MS imaging	Junior radiologist, multidisciplinary collaboration, technology curiosity

Table 2 - Phase 1 Participant Characteristics: Clinical and Technical Experts in MS Imaging and AI Development

All audio recordings were manually transcribed and cleaned by the research team. The 6th author, who is bilingual in English and French, conducted the initial thematic analysis across both language datasets, preserving original meaning and clinical terminology.

3.3 Phase 2: Expanding the Stakeholder Ecosystem

Phase 1 revealed that XAI adoption extends beyond the clinical-technical span, involving ethical, legal, and design considerations that neither radiologists nor engineers could fully address. Phase 2 therefore expanded our inquiry to include 6 domain experts recruited through purposive sampling (Table 3): 2 designers (1 specializing in XAI and Interface design, one in democratic AI and participatory methods), 1 AI ethicist (focusing on structural injustice and technology ethics), 1 lawyer (expertise in medical law and data protection), 1 AI engineer/researcher (developing XAI methods), 1 annotation specialist (MS imaging segmentation). These participants, predominantly holding doctorates and professorships with 10-20+ years of experience, were affiliated with universities and university hospitals across Switzerland, the Netherlands, and the United States.

Interview protocols were semi-structured but adapted to each participant's disciplinary expertise. Core common topics included disciplinary definitions of trust in AI contexts, transparency requirements from their professional perspective, ethical and legal concerns specific to healthcare AI, design recommendations for XAI systems, and reflections on interdisciplinary collaboration challenges and opportunities. For designers, questions emphasized user agency, interface design, and participatory methods. For the ethicist, questions centered on structural justice, bias, and power dynamics. For the legal expert, questions addressed regulatory frameworks, liability, and data protection. For the annotator, questions explored data quality, ground truth establishment, and practical annotation challenges. This flexibility allowed experts to draw on their specialized knowledge while maintaining comparability across interviews.

As with Phase 1, interviews were conducted in English and remotely via Zoom, recorded with consent, and lasted 45-60 minutes. The same transcription process was followed to preserve authenticity and meaning.

ID	Role	Specialization	Key details
S1	Ethicist	AI & tech ethics, critical theory, structural injustice	Focus: bias, data representation, trust complexity
S2	Designer	XAI archiving, participatory methods, socio-political AI issues	Focus: process visibility, training, system understanding
S3	Annotator	MS segmentation for Swiss MS Cohort	Focus: data labeling, practical annotation challenges
S4	Designer	Contestable AI, AI in public sector	Focus: user agency, autonomy, control in AI systems
S5	Engineer	AI-assisted medical image analysis tools	Focus: auditing mechanisms, patient safety, preventing over-reliance
S6	Legal Expert	Medical law, health law, data protection	Focus: regulatory trust, legal frameworks

Table 3 - Phase 2 Participant Characteristics: Domain Experts from Design, Ethics, Law, and Annotation

Phase 2 interviews were conducted on Zoom, transcribed, and refined using ChatGPT 4o to group responses by question and improve readability. Transcripts were verified against audio recordings for accuracy before importing into NVivo for inductive thematic analysis (60).

3.4 Phase 3: Patient Perspectives Through Participatory Workshop

Patients represent those most impacted by healthcare AI yet remain largely excluded from development processes. To address this gap, we designed a participatory workshop titled "Our Vulnerable Bodies: AI in Healthcare". The workshop engaged 6 participants (WS1-WS6, aged 20-25, 3 men and 3 women, all with design or engineering educational backgrounds) as proxies for patient perspectives. These individuals were recruited through informal networks including social media and word-of-mouth, ensuring multidisciplinary representation while acknowledging the limitations of using healthy participants rather than actual MS patients.

The 90-minute workshop, conducted in-person, was facilitated by the first two authors - one leading activities while the other documented observations. The workshop progressed through five carefully designed activities:

- First, an icebreaker exercise (15 minutes) was inspired by the New Metaphors method (61). It invited participants to select picture cards representing "bodily vulnerability" and share their interpretations, establishing a foundation for discussing healthcare experiences.

- Next, a sense-making activity (15 minutes) asked participants to connect their metaphors to AI in healthcare using provided keyword cards (e.g., "trust," "algorithmic bias," "power relations," "consent or dissent," "invisible boundaries") or writing their own. Discussion bridged embodied vulnerability and socio-technical concepts, surfacing concerns not articulated through direct questioning.
- The central material-making activity (30 minutes) challenged pairs of participants to create a wearable "armor" symbolizing protection in AI-assisted healthcare. To create the armors, they had to use simple provided materials like cardboard, fabrics, tape, scissors, and markers. This embodied design exercise surfaced tacit knowledge about vulnerability, safety, and agency that verbal interviews often cannot access. The constraints in materials and wearability were determined to encourage intuitive, emotionally-driven choices revealing underlying concerns.
- Following creation, groups participated in presentation and discussion (25 minutes) in a two-stage format. First, pairs interpreted others' creations before their own creators explained them, revealing independently legible meanings. In the second stage, creators presented their actual design intentions. Comparing projected versus authorial interpretations revealed shared understandings and distinctive meanings and fostered a common discussion.
- Finally, a reflection moment (5 minutes) gathered personal and collective reflections on the workshop experience.

Data collection included field notes (verbal exchanges, body language, group dynamics), photographs (setup, process, artifacts), and physical prototypes retained as material data. No audio recordings were kept to reduce performance anxiety. First author conducted interpretive analysis synthesizing patterns through team discussions, attending to emotional themes, metaphorical reasoning, and embodied vulnerability/protection expressions.

4 Results

We synthesized findings across all three phases through iterative team discussions, comparing insights across stakeholder groups and methods to identify convergences, divergences and unique contributions from each phase. This cross-phase integration through merging (62) generated a comprehensive, multi-perspective of stakeholder perspectives. Results are presented thematically, with themes emerging inductively from our empirical analysis, and are summarized in Table 4.

THEME	KEY INSIGHTS	STAKEHOLDER PERSPECTIVES
4.1 Trust as Critical Balance	<ul style="list-style-type: none"> Trust conceptualized fundamentally differently across stakeholder groups Trust requires calibration, not maximization Dual-sided risk: 	<p>CLINICIANS : Framed trust as SKEPTICISM and VERIFICATION; positioned AI as second opinion requiring verification; distinguished junior/senior needs; noted time pressure increases over-reliance risk</p> <p>ENGINEERS: Framed trust as TECHNICAL ACCOUNTABILITY;</p>

	<p>undertrust AND over-reliance both threaten safety</p> <ul style="list-style-type: none"> • Trust needs vary by expertise level and adoption stage 	<p>Focused on technical accountability and audit mechanisms; acknowledged limited clinical workflow understanding; recognized this as barrier to appropriate AI development</p> <p>DOMAIN EXPERTS: Framed trust as CONTEXT-DEPENDENT COMPLEXITY; proposed friction as design strategy; emphasized autonomy preservation; warned against both over-reliance and undertrust</p>
<p>4.2 Workflow and Audience</p> <p>Context-Dependence in the case of Multiple Sclerosis</p>	<ul style="list-style-type: none"> • MS diagnosis vs. follow-up require fundamentally different AI support • Diagnosis: multifactorial, resistant to automation • Follow-up: pattern-matching but verification still required • Lesion load variability affects AI 	<p>CLINICIANS: noted diagnosis requires multifactorial integration (imaging, labs, clinical presentation, McDonald criteria); described follow-up as repetitive; noted lesion load variability affects AI utility - high loads benefit from AI safety net, low loads risk cognitive burden from false positives; expressed concern that AI "doesn't save time" if verification required</p> <p>ENGINEERS: Focused on technical challenges; noted users unaware of</p>

	utility (high loads benefit, low loads add burden)	algorithm changes which undermines trust
4.3 Divergent Transparency Requirements Across Stakeholders	<ul style="list-style-type: none"> • No overlap in transparency needs across stakeholder groups • Each group requires fundamentally different information • Technical sophistication meaningless without deployment context fit 	<p>CLINICIANS (via Engineers): Need lesion count and location, not algorithmic details; prioritize "what" over "how"</p> <p>ENGINEERS: Need uncertainty quantification, performance metrics, validation; requirements share no overlap with clinical needs</p> <p>DOMAIN EXPERTS: Prioritized accountability mechanisms, dataset provenance, bias detection, contestability; emphasized training on risks and system limitations; positioned designers as "translators" needing multimodal explanations</p> <p>PATIENTS: Focused on personal impact rather than algorithmic mechanics; connected transparency to "consent/dissent" and "power relations"</p>

<p>4.4 Patient Perspectives on AI and Clinical Opacity</p>	<ul style="list-style-type: none"> • Patients face two opacity layers: clinical reasoning AND AI • AI adds second layer of complexity rather than creating new opacity • Data provenance matters more than algorithmic explanations 	<p>PATIENTS: Participants identified doctors as existing black boxes; emphasized medical terminology creates comprehension gaps independent of AI; all pairs created "armor" designs with windows/openings expressing desire to see into systems while seeking protection</p> <p>CLINICIANS: focus on diagnostic accuracy without mentioning patient comprehension; framed communication limits as time constraints; workflow design centers professional needs</p> <p>DOMAIN EXPERTS: Emphasized emotional, not just cognitive accessibility; shifted focus from algorithms to dataset transparency and diverse population representation; highlighted how AI risks perpetuating biases and power imbalances</p>
---	--	---

<p>4.5 Embodied Vulnerability</p>	<ul style="list-style-type: none"> • Emotional safety dominates patient concerns but absent from professional discourse • Consent experienced as emotional labor with unclear boundaries • Critical gap between cognitive focus of AI evaluation and emotional aspects of patient experience 	<p>PATIENTS: All selected vulnerability-themed cards; articulated consent as emotional labor; created wearable "armor" emphasizing privacy and emotional safety; no armor represented concepts related to diagnostic accuracy</p> <p>CLINICIANS: Only one brief mention of patient emotional states (qualified by time constraints); focused on diagnostic accuracy and workflow efficiency</p> <p>ENGINEERS: Expressed no awareness of emotional dimensions.</p> <p>DOMAIN EXPERTS: Emphasized "making sure people still feel in control"; advocated involving impacted communities early; framed as control rather than emotional safety</p>
--	---	--

Table 4 - Synthesis of Results

4.1 Trust as Critical Balance

All stakeholder groups emphasized trust as fundamental to AI adoption, yet conceptualizations diverged substantially from conventional technical approaches

that frame trust-building as the primary goal. Instead, participants across methods articulated trust as requiring careful calibration to match system capabilities, revealing a dual-sided risk: both insufficient trust (undertrust) and excessive trust (over-reliance) threaten safe AI integration.

Clinicians consistently emphasized the need for skepticism rather than confidence. As stated by C4, radiologists should maintain verification practices: "It's a bit like a check. We do it and the machine does the same thing and then we see if we agree". This comparative verification approach was echoed by five of six clinicians, which positioned AI as a second opinion requiring human control rather than a primary decision source. C6 articulated the verification requirement most explicitly, warning against automation that bypasses human judgment, and C5 cautioned: "I think we shouldn't skip too many steps in the review and interpretation phase", revealing deep concerns about automation bias across four clinical interviews.

The risk of over-reliance manifests acutely under time pressure. E1 observed that clinicians "won't give a second look to the decisions of AI model. They don't have time, and they make decisions based on it." E3 noted this creates a dangerous dynamic: "If we show algorithm decisions, it will be very difficult for a doctor to change their mind," highlighting the need for "systematic verification methods" to counteract automation bias. Conversely, C5 projected that "confidence level will increase with algorithm improvement," suggesting future risk of fully automated reporting without radiologist verification. Four clinicians (C1, C2, C3, C5) distinguished junior and senior needs, noting juniors might benefit from AI as

learning scaffolding while facing greater over-reliance risks, with C2 openly suggesting that appropriate trust levels must adapt to user expertise.

Trust calibration also follows a critical temporal trajectory. E5 described a vulnerable initial adoption phase: "If you lose the clinicians in this early phase, they'll never trust again. You really need to build a certain confidence." This contrasts with steady-state requirements for maintaining appropriate skepticism, suggesting different trust challenges at different adoption stages. Engineers also demonstrated awareness of reliability concerns. S5 articulated: "As an engineer, it's my responsibility to find a mechanism to audit that technology," framing trust as requiring technical accountability. However, E1 acknowledged their limited understanding of the system's final use case, expressing desire to observe the complete clinical workflow and see how clinicians manage patient cases. This reveals engineers' recognition of knowledge gaps regarding actual clinical practice, which four of five engineers mentioned as barriers to appropriate AI development.

Domain experts reframed trust as inherently complex and context-dependent. S1 challenged binary conceptualizations: "I think it's definitely normal and healthy for people to decide what to trust and to what degree they want to trust something". This view positions skepticism as desirable for appropriate trust calibration. S2 proposed intentional friction as design strategy: "It's a trade-off of understanding. The medical doctor needs to trust not only the result but also the process," arguing seamless integration might undermine appropriate skepticism. S1 elaborated this friction concept on an ethical perspective: "Having some friction or alerts to the users of the system," suggesting interface design should actively

promote critical engagement rather than passive acceptance of the explanations presented by the system. S4 emphasized preserving agency: "These systems have far-reaching consequences for people's autonomy (...) making sure people still feel in control of their lives while interacting with these AI systems." This connects trust calibration to autonomy preservation, suggesting interface design must enable, not eliminate, human judgment. S5 captured miscalibration risk: "Doctors don't even think about the result and just blindly trust the system. Bad things can happen." Conversely, four clinicians noted excessive caution could lead to undertrust, where clinicians dismiss helpful insights.

4.2 Workflow and Audience Context-Dependence in the case of Multiple Sclerosis

As discussed in 4.1, verification requirements and trust calibration challenges vary across contexts. The MS imaging workflow reveals how specific clinical tasks fundamentally shape both AI utility and transparency needs.

Clinicians distinguished sharply between diagnosis and follow-up scenarios, revealing task-specific requirements. C5 described diagnosis as inherently multifactorial: "There's a lot of information in an examination that is not necessarily related to the pathology for which patients are being followed, and so our job is also to look at everything else". Three clinicians (C3, C4, C5) noted imaging alone proves insufficient for diagnosis, requiring integration with laboratory results, clinical presentation, and patient history according to McDonald criteria. C1 elaborated that diagnosis requires proving "dissemination in space and time" across four lesion locations: periventricular, infratentorial, juxtacortical/cortical, and

spinal cord. This multifactorial diagnostic framework - requiring both spatial and temporal evidence - poses challenges for straightforward automation.

Follow-up monitoring presents different opportunities and constraints. Four clinicians (C2, C3, C4, C5) described this as repetitive "spot the difference" comparison - tracking lesion appearance, growth, and treatment response over time. This longitudinal, pattern-matching task appears more adaptable to AI assistance. However, a critical tension emerged: tasks where AI seems most applicable may not reduce workload if clinical responsibility mandates verification regardless of AI confidence. Notably, C4 cautioned: "it doesn't save time either" if verification remains required. Yet clinicians expressed willingness to verify AI outputs when systems provide capabilities exceeding human capabilities rather than merely automating existing ones. C6 noted that "We have an impression that we don't see everything and so we're limited.", so verification burden becomes acceptable when "the machine is more sensitive than our eye." However, S5 emphasized that quantifying this enhanced detection alone may be insufficient: "Guidance on consequences of decisions (for clinicians) I think are more important than how sure the system is."

This tension intensifies with lesion load variability. C6 noted that with high lesion loads, AI "can detect things we have trouble seeing with the eye" - acting as a safety net during lengthy slice-by-slice review. Conversely, C2 observed that in low lesion load cases, AI might add cognitive burden by highlighting false positives requiring investigation, "potentially slowing rather than accelerating workflow." Each AI-flagged candidate, whether true or false positive, demands clinical attention to verify or dismiss. Technical robustness compounds these

challenges. C5 noted scanner variability: "Depending on machines, magnetic field, brand - there could be false positives or negatives, so evaluation stability is crucial." E3 identified a related but distinct challenge in algorithm updates and retraining create unpredictability for clinicians: "Users aren't aware the algorithm changes, but unconsciously it influences confidence - not only do I not understand its internal rules, but it's something that moves". E3 elaborated that retraining introduces unpredictable changes in algorithm behavior, preventing users from developing stable expectations of the system.

4.3 Divergent Transparency Requirements Across Stakeholders

Beyond workflow dependencies, transparency requirements vary even more dramatically across stakeholder audiences than across clinical tasks. Engineers claimed clinicians prioritized understanding *what* AI detected rather than *how* algorithms function. E1 stated: "I don't think that clinicians want to have an understanding for each prediction. (...) The AI function is complex, so the explanation will be complex and they would spend 10 times more just trying to analyse why." E2 corroborated this, noting clinicians need "the number of lesions and the position of the lesion inside the brain" rather than algorithmic details.

Engineers required entirely different transparency, as emerged during the interviews: E1 focused on uncertainty quantification, E2 on performance metrics, E3 on algorithm validation. These engineering needs share no overlap with clinical needs. Domain experts introduced additional dimensions: S6 prioritized accountability when errors occur; S1 emphasized dataset provenance and bias detection; S4 connected transparency to contestability. S6 articulated training and

transparency as intertwined: "Medical doctors also need to be trained on the risks and on what can happen. Proper information should be provided, including, for example, how it was trained and what population it was trained on." This suggests transparency extends beyond interface design to encompass education about system provenance, limitations, and appropriate use contexts.

For workshop participants, transparency takes again a different meaning: understanding how decisions affect them personally rather than how algorithms work abstractly. During *sense-making*, participants connected transparency to "consent or dissent" and "power relations between people," revealing concerns about data usage and decision impact. Material-making produced designs of armors that bore no resemblance to technical explanations like saliency maps.

In this complex context, S2 positioned designers as translators requiring "multimodal explanations," yet no participant described concrete orchestration methods. The pattern is consistent across participants: XAI utility emerges from fit with specific contexts rather than intrinsic algorithmic properties. Transparency requirements diverge fundamentally across audiences with no overlap between professional and patient needs. Technical sophistication means little if deployment contexts are misunderstood and explanations serve only narrow stakeholder groups.

4.4 Patient Perspectives on AI and Clinical Opacity

Workshop participants revealed that both AI systems and clinicians operate as "black boxes" from patient perspectives, appearing consistently across all six participants. During the sense-making activity, participants connected vocabulary

cards including "algorithmic bias," "invisible boundaries," and "trust" to their selected metaphor cards representing bodily vulnerability. One participant explicitly stated: "Doctors are also black boxes - they give you information you cannot fully understand. AI feels the same unless it is explained." Another noted: "There is also a difference in vocabulary between you and a doctor," highlighting how medical terminology creates comprehension gaps independent of AI complexity.

AI intensifies this burden with data protection questions and algorithmic accountability concerns. Clinical interviews corroborated communication limitations though framed as time constraints. C3 described time pressure emphasizing efficiency over patient understanding. Four of six clinicians focused on diagnostic accuracy without mentioning patient comprehension, suggesting workflow design centers professional needs while marginalizing patient information needs.

Recognizing this multi-stakeholder complexity, domain experts recognized that transparent AI solutions must address various dimensions simultaneously. S2 emphasized explanations must address "emotional accessibility, not just cognitive" highlighting that transparency requirements extend beyond making information available to ensuring it is psychologically processable for diverse stakeholders. S1 reframed transparency entirely: "Trust starts with transparency in the datasets, especially how they represent diverse populations," shifting focus from algorithmic explanations to data provenance. This reframing connects technical transparency to broader questions of representational justice and epistemic equity in AI development. They further elaborated on how AI systems

risk perpetuating "biases and power imbalances inherent in current medical practices, thereby exacerbating inequities at a socio-technical level," suggesting transparency requirements extend beyond technical explanations to address systemic justice issues embedded in training data and deployment contexts.

4.5 Embodied Vulnerability

Workshop participants foregrounded emotional dimensions when discussing AI in healthcare. Throughout their responses, participants emphasized the concept of emotional safety - understood here as a sense of feeling secure and taken care of beyond the absence of physical harm (63). All six participants selected photographic cards representing vulnerability during icebreaker. One explained: "Sometimes you feel protected in a safe environment to talk about something that makes you feel involved". Material-making produced tangible evidence of these concerns: all three pairs created wearable "armor" designs symbolizing privacy, emotional safety, and control. Their designs always incorporated windows, openings, and transparent elements. One pair described their armor as representing "privacy and emotional safety - something I want protected when dealing with AI in healthcare." None of the armors' designs referenced diagnostic accuracy or clinical effectiveness - dimensions dominating professional discourse - instead materializing emotional and relational concerns in a physical form.

Five of six participants articulated consent as emotional labor: evaluating disclosure risks, anticipating data usage, managing fears of judgment, and maintaining dignity. In the sense-making activity, keyword selections revealed underlying concerns: "power relations between people," "marginalized bodies,"

"invisible boundaries," and "consent or dissent" appeared across all six participants, suggesting shared anxieties about surveillance, loss of control, and being reduced to data points. This concern was also reflected in the interviews, where S1 emphasized the importance of involving those most impacted: "I think designers or researchers should involve communities their product might impact. By engaging stakeholders early on, they can identify issues and concerns that might otherwise be missed," connecting emotional safety to participatory design processes. S4 also emphasized the importance of "making sure people still feel in control". In contrast, clinicians emphasized professional responsibility, verification requirements, and avoiding automation bias, framing their concerns primarily in terms of workflow integrity and clinical thoroughness rather than explicitly about patients' emotional states. Only C1 briefly noted communication importance but qualified with time constraints. No engineers mentioned emotional dimensions across five interviews.

5 Discussion

XAI utility and transparency requirements depend fundamentally on clinical contexts and stakeholder positions, challenging technical development assumptions of universal applicability.

5.1 Beyond Trust Calibration: Divergent Frameworks Across Stakeholders

Our findings challenge a central assumption underlying technical XAI development: **trust is not merely something to calibrate but is conceptualized fundamentally differently across stakeholder groups**. While existing research

establishes that users should neither overtrust nor undertrust AI systems (21,22,64,65), our inquiry reveals that what constitutes "trust" itself varies so substantially across professional communities that designing for "calibrated trust" requires first reconciling incommensurable framings.

Clinicians framed trust as skepticism and verification practice, **emphasizing AI as second opinion requiring human evaluation rather than primary decision source.**

This framing positions appropriate trust as maintaining professional doubt rather than building confidence, conflicting with narratives that position efficiency as a driver for AI adoption. If questioning or challenging AI assessments inevitably extends the time required for diagnosis, then clinicians cannot simultaneously save time and maintain the verification practices they identify as essential for appropriate trust.

Engineers framed trust as technical accountability and audit mechanisms, emphasizing systematic validation rather than interpersonal verification. Engineers focused on performance metrics, uncertainty quantification, and algorithm reliability - dimensions largely absent from clinical discourse. Critically, engineers acknowledged limited clinical workflow understanding, constraining their ability to design for trust as clinicians experience it.

Domain experts framed trust as context-dependent complexity requiring preservation of critical agency. S1 challenged binary conceptualizations entirely, describing trust as inherently complex and emphasizing that deciding what to trust and to what degree represents healthy skepticism rather than a problem to solve. This fundamentally contradicts technical approaches treating user doubt as a

barrier to overcome. S2 proposed intentional friction as design strategy, arguing that seamless integration might undermine appropriate critical engagement. S4 connected trust to autonomy preservation, framing it as maintaining human control and agency rather than building confidence in system capabilities.

Current XAI research typically assumes a shared understanding of trust that requires calibration to appropriate levels (21,66). Our findings suggest this idea overlooks a fundamental disagreement about what makes AI systems trustworthy and what relationships humans should have with algorithmic recommendations.

Supporting appropriate reliance on MS imaging XAI must therefore move beyond universal principles and adopt stakeholder-specific design approaches. This may mean verification-focused interfaces for clinicians reviewing lesion loads, accountability documentation for engineers validating algorithms, autonomy-preserving mechanisms for MS patients navigating treatment decisions. More fundamentally, transdisciplinary XAI development must begin with explicit dialogue about what trustworthiness means across professional communities, rather than assuming improved explanations will uniformly build trust.

5.2 MS Clinical Workflow and Regulatory Constraints

Our MS imaging findings reveal that **context-dependence operates not merely as design consideration but as multilevel structural constraint** (clinical, regulatory, and technical) that shapes adoption possibilities regardless of algorithmic sophistication or interface quality.

Clinicians distinguished sharply between MS diagnosis and follow-up monitoring, revealing how different clinical tasks fundamentally shape both AI utility and

adoption feasibility. MS diagnosis requires multifactorial integration resistant to straightforward automation, while follow-up monitoring - repetitive longitudinal comparison tracking lesion changes - appears more suitable for AI assistance. Yet even in suitable contexts, C4's observation that verification requirements prevent time-saving reveals a barrier. Professional liability and standards of care reinforce this requirement through legal mandate: radiologist oversight remains obligatory regardless of AI accuracy, compounding the tension between verification practices and efficiency narratives that typically justify clinical AI adoption. Technical infrastructure, like scanner variability and algorithm version instability, introduce additional variability.

AI-specific regulatory frameworks introduce additional constraints on deployment. Regulations like GDPR and the EU AI act mandate explainability for automated decision-making affecting individuals; medical device regulations require algorithmic auditability, performance documentation, and clinical safety evidence. These demands shape what transparency must accomplish: not only supporting clinical understanding but establishing legal accountability. S6 noted current frameworks inadequately address this accountability distribution between clinician, algorithm developer, and healthcare institution. Legal uncertainty thus operates as an adoption barrier independent of technical trustworthiness (67).

Critically, these constraints interact rather than operate independently. Clinical efficiency goals intersect with regulatory auditability demands; liability frameworks constrain what automation can achieve regardless of technical accuracy; infrastructure heterogeneity affects trust dynamics beyond interface control.

Recognizing these constraints explicitly is therefore essential for any realistic XAI integration strategy.

5.3 The Double Black Box

A critical finding that emerged exclusively through participatory methods challenges a foundational XAI assumption: that transparency problems originate with AI opacity. Previous work has documented information asymmetries between patients and clinicians in chronic illness management (68–70), yet XAI research overwhelmingly treats clinicians as sole end-users (38), assuming transparency stops at the professional boundary. Our work extends this by revealing how AI compounds existing communicative challenges rather than creating new ones.

Workshop participants consistently articulated that **both AI systems and clinicians operate as "black boxes" from a patient perspective**. We conceptualize this as the **double black box phenomenon**: from this experiential viewpoint, AI introduces a second opacity layer in patient-clinician relationships where the first layer - clinical expertise and medical reasoning - already creates comprehension challenges. While the double black box metaphor has been applied in other domains (71,72), we apply it here to reveal how AI compounds pre-existing patient-clinician information asymmetries. Notably, the clinicians we interviewed operate within a structural communication gap: radiologists rarely interact directly with MS patients, as neurologists typically serve as primary points of contact.

Workshop findings also revealed that patients prioritize different dimensions of transparency than professionals. Participants foregrounded emotional safety which is increasingly recognized as integral to patient safety itself, as its absence

can lead to fear, mistrust, medical trauma and reduced healthcare utilization (73). Their material designs incorporated protective barriers with transparent openings, structural features that expressed participants' dual needs to see into systems affecting them while maintaining protection, and their keyword selections revealed anxieties entirely absent from professional discourse focused on diagnostic accuracy. This systematic divergence reveals that current XAI evaluation approaches prioritizing cognitive clarity and technical performance fundamentally misalign with the emotional, relational, and embodied dimensions central to patient experience. However, as our findings emerged from lay patients rather than ones in established care relationships, it is possible that communicative opacity diminishes as patient-physician relationships develop over time. This could suggest XAI interventions might be most beneficial at different points in the care trajectory.

Theoretically, this repositions the explainability problem. Rather than treating XAI as a technical challenge of making AI interpretable to clinicians, the double black box reveals XAI as an infrastructure challenge requiring support for communication across multiple expertise layers. In this lens, AI in clinical contexts

- specifically for MS care, where patients undergo repeated imaging over decades
- creates new coordination requirements for patients: they must now integrate their embodied illness experience not only with clinical interpretation, but also AI-assisted findings.

Yet XAI design renders this patient work invisible, focusing solely on clinician-algorithm interaction.

These findings suggest XAI researchers must reconceptualize transparency as a relational process spanning entire MS care pathways, not merely technical

property of AI systems. Future work must prioritize direct collaboration with MS patients, acknowledging that such participation demands resources, institutional commitment, and methodological innovation that current research practices rarely provide.

5.4 Power and participation in XAI Development

Despite widespread claims of user-centeredness in "human-centered" XAI development (49), our investigation reveals persistent gaps between participatory rhetoric and practice. Our positioning as design researchers embedded within the MSxplain transdisciplinary consortium was essential for enabling this multi-stakeholder inquiry, allowing us to engage clinical, technical, and public perspectives within a shared research infrastructure. Each stakeholder group surfaced knowledge invisible to others - most strikingly, only workshop participants centered emotional safety and bodily autonomy, dimensions entirely absent from professional and technical discourse, demonstrating that creative participatory methods access forms of situated knowledge that structured interviews cannot reach (74).

Patient and Public Involvement (PPI) frameworks also emphasise meaningful engagement with those who have lived experience of conditions under study (75), yet achieving genuine participation with those most affected proved structurally difficult. Ethics review processes designed to protect vulnerable populations simultaneously created exclusion in participation pathways, meaning that despite institutional support we could not include MS patients directly. Moving beyond performative participation requires concrete transformation: review processes

accommodating sustained engagement, funding models supporting longitudinal participatory research, and recognition that genuine inclusion demands timeline adjustments and resources that current research economies rarely provide.

5.5 Design Implications for MS Imaging XAI

Our findings yield specific design implications that challenge conventional XAI development assumptions, requiring approaches that acknowledge meaning divergence, multilevel constraints, and the double black box phenomenon.

Supporting heterogeneous trust framework. As trust operates differently across stakeholder groups, requiring systems that support multiple frameworks simultaneously rather than optimizing for a single conceptualization. For clinicians, interfaces should present AI-detected lesions alongside radiologist workspace for side-by-side comparison, positioning AI as second opinion requiring verification rather than prescriptive recommendation. For engineers, systems must provide accessible audit trails, performance metrics, and uncertainty quantification enabling technical accountability. For domain experts concerned with structural justice, systems should document dataset provenance, demographic representation, and known bias patterns.

Dynamic context adaptation. Rather than presenting identical explanations across cases, systems should adjust information depth and interaction patterns based on clinical context and AI confidence. For high-uncertainty lesion detections, interfaces should provide detailed uncertainty communication and comparative visualizations requiring active verification, implementing the "intentional friction" domain experts proposed to prevent over-reliance. For high-certainty follow-up

monitoring with clear lesion boundaries, streamlined outputs can reduce cognitive burden while maintaining awareness of limitations. System representations should adapt across uncertainty levels, task type (diagnosis versus follow-up), lesion load characteristics, scan quality, and user expertise. Critically, regulatory information, like algorithm version, training data characteristics, known limitations, performance benchmarks, must remain continuously accessible, embedding compliance within interaction patterns rather than separate documentation. Clinician training must emphasize system limitations, failure modes, and appropriate reliance strategies across MS imaging contexts.

Extending beyond clinician-facing transparency. The double black box phenomenon presents a design opportunity for digital health practitioners to reframe transparency: rather than simply "opening" algorithmic black boxes (76), the challenge becomes negotiating and operationalizing understanding through interaction design, making both AI and clinical reasoning meaningfully graspable for patients. XAI systems could bridge existing communicative gaps by addressing both layers of opacity simultaneously. Potential approaches include generating plain-language summaries clinicians can share with MS patients during consultations, or developing patient-accessible interfaces showing how AI findings contributed to clinical assessments.

These implications require iterative prototyping and evaluation with radiologists across diverse MS imaging scenarios. Implementation demands institutional infrastructure supporting differentiated stakeholder access, regulatory compliance, and comprehensive training programs, not merely interface refinement.

5.6 Limitations

This study's limitations reflect both methodological choices inherent to qualitative inquiry and structural barriers encountered during transdisciplinary research. Our interpretive approach prioritized depth over breadth, investigating 23 selected participants to surface rich, contextualized insights that quantitative methods alone could not capture. Most significantly, we employed healthy proxies rather than actual MS patients in our participatory workshop. This limitation reveals structural challenges facing participatory XAI development: despite institutional support, ethical approval, and research resources, we could not overcome access barriers to genuinely include those most affected. Additional limitations include geographic concentration (Swiss healthcare context), temporal scope (perspectives during early development, not post-deployment), and sample size limiting demographic diversity.

Future research should address these gaps through sustained co-design with MS patient communities, longitudinal studies tracking trust dynamics through deployment phases and mixed-methods approaches combining our qualitative insights with quantitative validation of the design implications we propose. Critically, the design strategies articulated in Section 5.5, particularly dynamic adaptive representations adjusting to uncertainty and context, require iterative prototyping and evaluation with radiologists across diverse MS imaging scenarios.

Conclusions

This transdisciplinary investigation of explainable AI for Multiple Sclerosis imaging reveals fundamental misalignments between technical XAI development assumptions and the heterogeneous requirements of clinical implementation.

Through three-phase generative research engaging radiologists, engineers, domain experts, and lay public participants, we uncovered empirical insights that challenge prevailing assumptions in technical XAI development.

Our findings reveal that trust is conceptualized fundamentally differently across stakeholder groups, requiring reconciliation rather than universal solutions. The "double black box" phenomenon proved particularly striking: workshop participants recognized that both AI systems and clinicians operate opaquely from patient perspectives, suggesting transparency challenges extend beyond algorithmic explainability to encompass entire clinical communication practices.

The MS imaging workflows expose how context-dependence operates as multilevel structural constraints (clinical, regulatory, technical) that shape adoption possibilities independently of algorithmic sophistication or interface quality.

From these insights, we identify design implications that foreground stakeholder diversity, requiring XAI interfaces to support multiple trust calibration strategies simultaneously, facilitate transparency across professional and patient-facing contexts, and privilege emotional safety alongside technical performance.

This work demonstrates design research's capacity to surface knowledge invisible in single-stakeholder or purely computational approaches, particularly regarding emotional and embodied dimensions of vulnerability. For AI-assisted MS imaging

specifically, successful clinical integration requires addressing workflow constraints, regulatory frameworks, and multilayered communication needs simultaneously. Future research must prioritize sustained transdisciplinary dialogue, co-design with patient communities, developing institutional infrastructures that enable genuine participation rather than performative consultation. Only through such fundamental reorientation can XAI meaningfully support trust, clinical expertise, and patient autonomy in MS care.

Acknowledgements

We thank the MSxplain transdisciplinary research consortium for their continuous support throughout this research. We are grateful to all participants who contributed to this study. We also extend our thanks to Kars Alfrink and Ting-an Lin for their valuable input and participation.

References

1. Reyes M, Meier R, Pereira S, Silva CA, Dahlweid FM, Tengg-Kobligk HV, et al. On the Interpretability of Artificial Intelligence in Radiology: Challenges and Opportunities. *Radiol Artif Intell*. 2020 May 1;2(3):e190043. doi:10.1148/ryai.2020190043
2. Haas S, Hegestweiler K, Rapp M, Muschalik M, Hüllermeier E. Stakeholder-centric explanations for black-box decisions: an XAI process model and its application to automotive goodwill assessments. *Front Artif Intell*. 2024 Oct 24;7:1471208. doi:10.3389/frai.2024.1471208
3. Borys K, Schmitt YA, Nauta M, Seifert C, Krämer N, Friedrich CM, et al. Explainable AI in medical imaging: An overview for clinical practitioners – Beyond saliency-based XAI approaches. *Eur J Radiol*. 2023 May;162:110786. doi:10.1016/j.ejrad.2023.110786
4. Abrantes J, Rouzrokh P. Explaining explainability: The role of XAI in medical imaging. *Eur J Radiol*. 2024 Apr;173:111389. doi:10.1016/j.ejrad.2024.111389

5. Chaddad A, Peng J, Xu J, Bouridane A. Survey of Explainable AI Techniques in Healthcare. *Sensors*. 2023 Jan 5;23(2):634. doi:10.3390/s23020634
6. Walton C, King R, Rechtman L, Kaye W, Leray E, Marrie RA, et al. Rising prevalence of multiple sclerosis worldwide: Insights from the Atlas of MS, third edition. *Mult Scler J*. 2020 Dec;26(14):1816–21. doi:10.1177/1352458520970841
7. Janssens ACJW, Van Doorn PA, De Boer JB, Van Der Meché FGA, Passchier J, Hintzen RQ. Impact of recently diagnosed multiple sclerosis on quality of life, anxiety, depression and distress of patients and partners: **Quality of life and emotional well-being in MS**. *Acta Neurol Scand*. 2003 Dec;108(6):389–95. doi:10.1034/j.1600-0404.2003.00166.x
8. Topcu G, Mhizha-Murira JR, Griffiths H, Bale C, Drummond A, Fitzsimmons D, et al. Experiences of receiving a diagnosis of multiple sclerosis: a meta-synthesis of qualitative studies. *Disabil Rehabil*. 2023 Feb 27;45(5):772–83. doi:10.1080/09638288.2022.2046187
9. Montalban X, Lebrun-Fréney C, Oh J, Arrambide G, Moccia M, Pia Amato M, et al. Diagnosis of multiple sclerosis: 2024 revisions of the McDonald criteria. *Lancet Neurol*. 2025 Oct;24(10):850–65. doi:10.1016/S1474-4422(25)00270-4
10. Lladó X, Oliver A, Cabezas M, Freixenet J, Vilanova JC, Quiles A, et al. Segmentation of multiple sclerosis lesions in brain MRI: A review of automated approaches. *Inf Sci*. 2012 Mar;186(1):164–85. doi:10.1016/j.ins.2011.10.011
11. Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Med Educ*. 2023 Sep 22;23(1):689. doi:10.1186/s12909-023-04698-z
12. Vereschak O, Bailly G, Caramiaux B. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proc ACM Hum-Comput Interact*. 2021 Oct 13;5(CSCW2):1–39. doi:10.1145/3476068
13. Seoni S, Jahmunah V, Salvi M, Barua PD, Molinari F, Acharya UR. Application of uncertainty quantification to artificial intelligence in healthcare: A review of last decade (2013–2023). *Comput Biol Med*. 2023 Oct;165:107441. doi:10.1016/j.combiomed.2023.107441
14. Ihongbe IE, Fouad S, Mahmoud TF, Rajasekaran A, Bhatia B. Evaluating Explainable Artificial Intelligence (XAI) techniques in chest radiology imaging through a human-centered Lens. *PLOS ONE*. 2024 Oct 9;19(10):e0308758. doi:10.1371/journal.pone.0308758
15. Spagnolo F, Depeursinge A, Schädelin S, Akbulut A, Müller H, Barakovic M, et al. How far MS lesion detection and segmentation are integrated into the clinical workflow? A systematic review. *NeuroImage Clin*. 2023;39:103491. doi:10.1016/j.nicl.2023.103491

16. Camaradou JCL, Hogg HDJ. Commentary: Patient Perspectives on Artificial Intelligence; What have We Learned and How Should We Move Forward? *Adv Ther.* 2023 Jun;40(6):2563–72. doi:10.1007/s12325-023-02511-3
17. Göttgens I, Oertelt-Prigione S. The Application of Human-Centered Design Approaches in Health Research and Innovation: A Narrative Review of Current Practices. *JMIR MHealth UHealth.* 2021 Dec 6;9(12):e28102. doi:10.2196/28102
18. Nguyen M, Mougnot C. A systematic review of empirical studies on multidisciplinary design collaboration: Findings, methods, and challenges. *Des Stud.* 2022 Jul;81:101120. doi:10.1016/j.destud.2022.101120
19. Zhan W, Motta M, Baez-Lugo S, Henchoz N, Cuadra MB, Lemay DR. Explainable AI and Trust, Design Methodologies to Explore Patients' Perspective. In: 2025 IEEE 38th International Symposium on Computer-Based Medical Systems (CBMS). 2025. p. 512–3. doi:10.1109/CBMS65348.2025.00110
20. Alonso M, Astobiza AM, Ortega Lozano R. AI-mediated healthcare and trust. A trust-construct and trust-factor framework for empirical research. *Artif Intell Rev.* 2025 Aug 20;58(11):337. doi:10.1007/s10462-025-11306-7
21. Asan O, Bayrak AE, Choudhury A. Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians. *J Med Internet Res.* 2020 Jun 19;22(6):e15154. doi:10.2196/15154
22. Panigutti C, Beretta A, Fadda D, Giannotti F, Pedreschi D, Perotti A, et al. Co-design of Human-centered, Explainable AI for Clinical Decision Support. *ACM Trans Interact Intell Syst.* 2023 Dec 31;13(4):1–35. doi:10.1145/3587271
23. Jones C, Thornton J, Wyatt JC. Artificial intelligence and clinical decision support: clinicians' perspectives on trust, trustworthiness, and liability. *Med Law Rev.* 2023 Nov 1;31(4):501–20. doi:10.1093/medlaw/fwad013
24. González-Gonzalo C, Thee EF, Klaver CCW, Lee AY, Schlingemann RO, Tufail A, et al. Trustworthy AI: Closing the gap between development and integration of AI systems in ophthalmic practice. *Prog Retin Eye Res.* 2022 Sep;90:101034. doi:10.1016/j.preteyeres.2021.101034
25. Okamura K, Yamada S. Adaptive trust calibration for human-AI collaboration. Lv C, editor. *PLOS ONE.* 2020 Feb 21;15(2):e0229132. doi:10.1371/journal.pone.0229132
26. Zhang Y, Liao QV, Bellamy RKE. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency [Internet].* Barcelona Spain: ACM; 2020 [cited 2025 Oct 22]. p. 295–305. Available from: <https://dl.acm.org/doi/10.1145/3351095.3372852> doi:10.1145/3351095.3372852
27. Yang Q, Hao Y, Quan K, Yang S, Zhao Y, Kuleshov V, et al. Harnessing Biomedical Literature to Calibrate Clinicians' Trust in AI Decision Support

- Systems. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems [Internet]. Hamburg Germany: ACM; 2023 [cited 2025 Oct 16]. p. 1–14. Available from: <https://dl.acm.org/doi/10.1145/3544548.3581393> doi:10.1145/3544548.3581393
28. Nasarian E, Alizadehsani R, Acharya UR, Tsui KL. Designing interpretable ML system to enhance trust in healthcare: A systematic review to proposed responsible clinician-AI-collaboration framework. *Inf Fusion*. 2024 Aug;108:102412. doi:10.1016/j.inffus.2024.102412
 29. Sadeghi Z, Alizadehsani R, Cifci MA, Kausar S, Rehman R, Mahanta P, et al. A review of Explainable Artificial Intelligence in healthcare. *Comput Electr Eng*. 2024 Aug;118:109370. doi:10.1016/j.compeleceng.2024.109370
 30. Ozdemir O, Fatunmbi TO. Explainable AI (XAI) in Healthcare: Bridging the Gap between Accuracy and Interpretability. *J Sci Technol Eng Res*. 2024 Jun 30;2(1):32–44. doi:10.64206/0z78ev10
 31. Branley-Bell D, Whitworth R, Coventry L. User Trust and Understanding of Explainable AI: Exploring Algorithm Visualisations and User Biases. In: Kurosu M, editor. *Human-Computer Interaction. Human Values and Quality of Life* [Internet]. Cham: Springer International Publishing; 2020 [cited 2025 Nov 11]. p. 382–99. (Lecture Notes in Computer Science). Available from: https://link.springer.com/10.1007/978-3-030-49065-2_27 doi:10.1007/978-3-030-49065-2_27
 32. Wang D, Yang Q, Abdul A, Lim BY. Designing Theory-Driven User-Centric Explainable AI. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems [Internet]. Glasgow Scotland Uk: ACM; 2019 [cited 2025 Nov 11]. p. 1–15. Available from: <https://dl.acm.org/doi/10.1145/3290605.3300831> doi:10.1145/3290605.3300831
 33. Abdul A, Vermeulen J, Wang D, Lim BY, Kankanhalli M. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems [Internet]. Montreal QC Canada: ACM; 2018 [cited 2025 Oct 16]. p. 1–18. Available from: <https://dl.acm.org/doi/10.1145/3173574.3174156> doi:10.1145/3173574.3174156
 34. Gambetti A, Han Q, Shen H, Soares C. A Survey on Human-Centered Evaluation of Explainable AI Methods in Clinical Decision Support Systems [Internet]. *arXiv*; 2025 [cited 2025 Oct 18]. Available from: <https://arxiv.org/abs/2502.09849> doi:10.48550/ARXIV.2502.09849
 35. European Parliament. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence

- Act) (Text with EEA relevance). OJ L 2024/1689 [Internet]. 2024 Jun 13. Available from: <http://data.europa.eu/eli/reg/2024/1689/oj>
36. Di Martino F, Delmastro F. Explainable AI for clinical and remote health applications: a survey on tabular and time series data. *Artif Intell Rev.* 2023 Jun;56(6):5261–315. doi:10.1007/s10462-022-10304-3
 37. Langer M, Oster D, Speith T, Hermanns H, Kästner L, Schmidt E, et al. What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artif Intell.* 2021 Jul;296:103473. doi:10.1016/j.artint.2021.103473
 38. Bienefeld N, Boss JM, Lüthy R, Brodbeck D, Azzati J, Blaser M, et al. Solving the explainable AI conundrum by bridging clinicians' needs and developers' goals. *Npj Digit Med.* 2023 May 22;6(1):94. doi:10.1038/s41746-023-00837-4
 39. Tonekaboni S, Joshi S, McCradden MD, Goldenberg A. What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use [Internet]. arXiv; 2019 [cited 2025 Dec 15]. Available from: <http://arxiv.org/abs/1905.05134> doi:10.48550/arXiv.1905.05134
 40. Pahud De Mortanges A, Luo H, Shu SZ, Kamath A, Suter Y, Shelan M, et al. Orchestrating explainable artificial intelligence for multimodal and longitudinal data in medical imaging. *Npj Digit Med.* 2024 Jul 22;7(1):195. doi:10.1038/s41746-024-01190-w
 41. Jacobs M, He J, F. Pradier M, Lam B, Ahn AC, McCoy TH, et al. Designing AI for Trust and Collaboration in Time-Constrained Medical Decisions: A Sociotechnical Lens. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* [Internet]. Yokohama Japan: ACM; 2021 [cited 2025 Oct 16]. p. 1–14. Available from: <https://dl.acm.org/doi/10.1145/3411764.3445385> doi:10.1145/3411764.3445385
 42. Fogliato R, Chappidi S, Lungren M, Fisher P, Wilson D, Fitzke M, et al. Who Goes First? Influences of Human-AI Workflow on Decision Making in Clinical Imaging. In: *2022 ACM Conference on Fairness Accountability and Transparency* [Internet]. Seoul Republic of Korea: ACM; 2022 [cited 2025 Oct 22]. p. 1362–74. Available from: <https://dl.acm.org/doi/10.1145/3531146.3533193> doi:10.1145/3531146.3533193
 43. Amann J, Vetter D, Blomberg SN, Christensen HC, Coffee M, Gerke S, et al. To explain or not to explain?—Artificial intelligence explainability in clinical decision support systems. Lu HHS, editor. *PLOS Digit Health.* 2022 Feb 17;1(2):e0000016. doi:10.1371/journal.pdig.0000016
 44. Chen H, Gomez C, Huang CM, Unberath M. Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review. *Npj Digit Med.* 2022 Oct 19;5(1):156. doi:10.1038/s41746-022-00699-2

45. Schoonderwoerd TAJ, Jorritsma W, Neerincx MA, van den Bosch K. Human-centered XAI: Developing design patterns for explanations of clinical decision support systems. *Int J Hum-Comput Stud*. 2021 Oct 1;154:102684. doi:10.1016/j.ijhcs.2021.102684
46. Delgado F, Yang S, Madaio M, Yang Q. The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice. In: *Equity and Access in Algorithms, Mechanisms, and Optimization* [Internet]. Boston MA USA: ACM; 2023 [cited 2025 Oct 22]. p. 1–23. Available from: <https://dl.acm.org/doi/10.1145/3617694.3623261> doi:10.1145/3617694.3623261
47. Birhane A, Isaac W, Prabhakaran V, Diaz M, Elish MC, Gabriel I, et al. Power to the People? Opportunities and Challenges for Participatory AI. In: *Equity and Access in Algorithms Mechanisms and Optimization* [Internet]. Arlington VA USA: ACM; 2022 [cited 2025 Oct 22]. p. 1–8. Available from: <https://dl.acm.org/doi/10.1145/3551624.3555290> doi:10.1145/3551624.3555290
48. Hossain S, Ahmed SI. Towards a New Participatory Approach for Designing Artificial Intelligence and Data-Driven Technologies [Internet]. *arXiv*; 2021 [cited 2025 Oct 22]. Available from: <https://arxiv.org/abs/2104.04072> doi:10.48550/ARXIV.2104.04072
49. Subramanian HV, Canfield C, Shank DB. Designing explainable AI to improve human-AI team performance: A medical stakeholder-driven scoping review. *Artif Intell Med*. 2024 Mar;149:102780. doi:10.1016/j.artmed.2024.102780
50. Stogiannos N, Gillan C, Precht H, Reis CSD, Kumar A, O'Regan T, et al. A multidisciplinary team and multiagency approach for AI implementation: A commentary for medical imaging and radiotherapy key stakeholders. *J Med Imaging Radiat Sci*. 2024 Dec;55(4):101717. doi:10.1016/j.jmir.2024.101717
51. Graziani M, Dutkiewicz L, Calvaresi D, Amorim JP, Yordanova K, Vered M, et al. A global taxonomy of interpretable AI: unifying the terminology for the technical and social sciences. *Artif Intell Rev*. 2023 Apr;56(4):3473–504. doi:10.1007/s10462-022-10256-8
52. Sendra P. The ethics of co-design. *J Urban Des*. 2024 Jan 2;29(1):4–22. doi:10.1080/13574809.2023.2171856
53. Sloane M, Moss E, Awomolo O, Forlano L. Participation Is not a Design Fix for Machine Learning. In: *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* [Internet]. New York, NY, USA: Association for Computing Machinery; 2022 [cited 2025 Jan 6]. p. 1–6. (EAAMO '22). Available from: <https://dl.acm.org/doi/10.1145/3551624.3555285> doi:10.1145/3551624.3555285
54. Lin TA, Chen PHC. Artificial Intelligence in a Structurally Unjust Society. *Fem Philos Q*. 2022 Dec 21;8(3/4):3/4. doi:10.5206/fpq/2022.3/4.14191

55. Zytka D, J. Wisniewski P, Guha S, P. S. Baumer E, Lee MK. Participatory Design of AI Systems: Opportunities and Challenges Across Diverse Users, Relationships, and Application Domains. In: CHI Conference on Human Factors in Computing Systems Extended Abstracts [Internet]. New Orleans LA USA: ACM; 2022 [cited 2025 Oct 22]. p. 1–4. Available from: <https://dl.acm.org/doi/10.1145/3491101.3516506> doi:10.1145/3491101.3516506
56. Lysen F, Wyatt S. Refusing participation: hesitations about designing responsible patient engagement with artificial intelligence in healthcare. *J Responsible Innov.* 2024 Dec 31;11(1):2300161. doi:10.1080/23299460.2023.2300161
57. Hodson E, Svanda A, Dadashi N. Whom do we include and when? participatory design with vulnerable groups. *CoDesign.* 2023 Oct 2;19(4):269–86. doi:10.1080/15710882.2022.2160464
58. Hogg HDJ, Al-Zubaidy M, Technology Enhanced Macular Services Study Reference Group, Talks J, Denniston AK, Kelly CJ, et al. Stakeholder Perspectives of Clinical Artificial Intelligence Implementation: Systematic Review of Qualitative Evidence. *J Med Internet Res.* 2023 Jan 10;25:e39742. doi:10.2196/39742
59. Palinkas LA, Horwitz SM, Green CA, Wisdom JP, Duan N, Hoagwood K. Purposeful Sampling for Qualitative Data Collection and Analysis in Mixed Method Implementation Research. *Adm Policy Ment Health Ment Health Serv Res.* 2015 Sep;42(5):533–44. doi:10.1007/s10488-013-0528-y
60. Braun V, Clarke V. Using thematic analysis in psychology. *Qual Res Psychol.* 2006 Jan 1;3(2):77–101. doi:10.1191/1478088706qp063oa
61. Lockton D, Singh D, Sabnis S, Chou M, Foley S, Pantoja A. New Metaphors: A Workshop Method for Generating Ideas and Reframing Problems in Design and Beyond. In: Proceedings of the 2019 on Creativity and Cognition [Internet]. San Diego CA USA: ACM; 2019. p. 319–32. Available from: <https://dl.acm.org/doi/10.1145/3325480.3326570> doi:10.1145/3325480.3326570
62. Fetters MD, Curry LA, Creswell JW. Achieving Integration in Mixed Methods Designs—Principles and Practices. *Health Serv Res.* 2013 Dec;48(6pt2):2134–56. doi:10.1111/1475-6773.12117
63. Minartz P, Aumann CM, Vondeberg C, Kuske S. Feeling safe in the context of digitalization in healthcare: a scoping review. *Syst Rev.* 2024 Feb 8;13(1):62. doi:10.1186/s13643-024-02465-9
64. Jacovi A, Marasović A, Miller T, Goldberg Y. Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency [Internet]. Virtual Event Canada: ACM; 2021 [cited 2025 Nov 5]. p.

624–35. Available from: <https://dl.acm.org/doi/10.1145/3442188.3445923>
doi:10.1145/3442188.3445923

65. Wischnewski M, Krämer N, Müller E. Measuring and Understanding Trust Calibrations for Automated Systems: A Survey of the State-Of-The-Art and Future Directions. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems [Internet]. Hamburg Germany: ACM; 2023 [cited 2025 Nov 5]. p. 1–16. Available from: <https://dl.acm.org/doi/10.1145/3544548.3581197> doi:10.1145/3544548.3581197
66. Yang Q, Steinfeld A, Rosé C, Zimmerman J. Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems [Internet]. Honolulu HI USA: ACM; 2020 [cited 2024 Nov 8]. p. 1–13. Available from: <https://dl.acm.org/doi/10.1145/3313831.3376301> doi:10.1145/3313831.3376301
67. De Bruijn H, Warnier M, Janssen M. The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. Gov Inf Q. 2022 Apr;39(2):101666. doi:10.1016/j.giq.2021.101666
68. Ancker JS, Witteman HO, Hafeez B, Provencher T, Van De Graaf M, Wei E. The Invisible Work of Personal Health Information Management Among People With Multiple Chronic Conditions: Qualitative Interview Study Among Patients and Providers. J Med Internet Res. 2015 Jun 4;17(6):e137. doi:10.2196/jmir.4381
69. Pichon A, Schiffer K, Horan E, Massey B, Bakken S, Mamykina L, et al. Divided We Stand: The Collaborative Work of Patients and Providers in an Enigmatic Chronic Disease. Proc ACM Hum-Comput Interact. 2021 Jan 5;4(CSCW3):1–24. doi:10.1145/3434170
70. Samal L, Fu HN, Camara DS, Wang J, Bierman AS, Dorr DA. Health information technology to improve care for people with multiple chronic conditions. Health Serv Res. 2021 Oct;56(S1):1006–36. doi:10.1111/1475-6773.13860
71. Deeks AS. The Double Black Box. In: The Double Black Box [Internet]. 1st ed. Oxford University Press New York, NY; 2025 [cited 2026 Jan 26]. p. 106–32. Available from: <https://academic.oup.com/book/59551/chapter/502644647> doi:10.1093/9780197520932.003.0005
72. Wahlström M, Tammentie B, Salonen TT, Karvonen A. AI and the transformation of industrial work: Hybrid intelligence vs double-black box effect. Appl Ergon. 2024 Jul;118:104271. doi:10.1016/j.apergo.2024.104271
73. Lyndon A, Davis DA, Sharma AE, Scott KA. Emotional safety *is* patient safety. BMJ Qual Saf. 2023 Jul;32(7):369–72. doi:10.1136/bmjqs-2022-015573
74. Liao QV, Varshney KR. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences [Internet]. arXiv; 2021 [cited 2025 Nov 4]. Available from: <https://arxiv.org/abs/2110.10790> doi:10.48550/ARXIV.2110.10790

75. Staniszewska S, Brett J, Simera I, Seers K, Mockford C, Goodlad S, et al. GRIPP2 reporting checklists: tools to improve reporting of patient and public involvement in research. *BMJ*. 2017 Aug 2;j3453. doi:10.1136/bmj.j3453
76. Storey VC, Lukyanenko R, Maass W, Parsons J. Explainable AI. *Commun ACM*. 2022 Apr;65(4):27–9. doi:10.1145/3490699