UNIVERSITÄT AUGSBURG
Universitätsbibliothek

OPEN ACCESS
OPUS AUGSBURG

Enhancing trust and agency: integrating citizen perspectives into AI-assisted shared decision-making in medicine

Paula Ziethmann, Fabian Stieler, Stefanie Kranz Walter, Dennis Hartmann, Bernhard Bauer, Kerstin Schlögl-Flierl

Angaben zur Veröffentlichung / Publication details:

Ziethmann, Paula, Fabian Stieler, Stefanie Kranz Walter, Dennis Hartmann, Bernhard Bauer, and Kerstin Schlögl-Flierl. 2026. "Enhancing trust and agency: integrating citizen perspectives into AI-assisted shared decision-making in medicine." *AI & Society* 41(5): 5339–54. https://doi.org/10.1007/s00146-026-02906-0.

Nutzungsbedingungen / Terms of use:

CC BY 4.0

Dieses Dokument wird unter folgenden Bedingungen zur Verfügung gestellt: / This document is made available under these conditions:
CC-BY 4.0: Creative Commons: Namensnennung
Weitere Informationen finden Sie unter: / For more information see:
https://creativecommons.org/licenses/by/4.0/deed.de

AI & SOCIETY (2026) 41:5339–5354
https://doi.org/10.1007/s00146-026-02906-0

OPEN FORUM

Enhancing trust and agency: integrating citizen perspectives into AI-assisted shared decision-making in medicine

Paula Ziethmann[1,2] · Fabian Stieler[3] · Stefanie Kranz Walter[4] · Dennis Hartmann[3] · Bernhard Bauer[3,5] · Kerstin Schlögl-Flierl[3,5]

Received: 7 April 2025 / Accepted: 1 February 2026 / Published online: 13 February 2026
© The Author(s) 2026

Abstract
As artificial intelligence (AI) becomes increasingly integrated into clinical environments, questions of trust, transparency, and shared decision-making come to the fore. This article examines how public perspectives can influence the ethical and technical development of AI tools in medicine, drawing on empirical insights from an interdisciplinary project focused on developing AI to support the diagnosis and treatment of skin cancer. Rather than treating ethical concerns as external to technical design, we argue that they must be addressed from within the development process. In our case, this was achieved by integrating citizen feedback into iterative design loops within our interdisciplinary team, fostering closer alignment between AI functionalities and public values. Through focus group discussions with citizens and a constructivist grounded theory approach, we identified three key areas of concern: the evolving doctor–patient relationship, patient agency in AI-supported care, and the influence of specific medical contexts on public evaluations of AI. This article illustrates how these citizen perspectives can be meaningfully connected with the medical and technical considerations shaping the development of AI.

Keywords AI in healthcare · Shared decision-making · Public trust · Participatory ethics · Transparency · Patient-centered AI

1 Introduction

The integration of artificial intelligence (AI) into medical practice promises far-reaching changes in how diseases are diagnosed and treated—including basal cell carcinoma (BCC), the most common form of skin cancer worldwide. Early detection is key to avoiding invasive procedures such as micro-controlled surgery. Yet, in clinical reality, BCC is often diagnosed only at later stages. AI-supported diagnostic tools could shift this paradigm by enabling earlier detection and thus more targeted, less invasive treatment approaches.

The OCTOLAB project addresses this potential by combining optical coherence tomography (OCT) for BCC *diagnosis* with a long-pulsed (1064 nm Nd:YAG) laser for BCC *treatment*. The project's AI system is trained to analyze OCT images to identify clinically relevant features, such as tumor thickness, boundary delineation, and histopathological subtype. Through image processing and segmentation techniques, tumor areas can be differentiated from surrounding healthy tissue—forming the basis for a personalized calibration of the laser parameters, including energy density, pulse duration, and repetition rate. This technical system is

✉ Paula Ziethmann
paula.ziethmann@iab.de

Fabian Stieler
fabian.stieler@uni-a.de

Stefanie Kranz Walter
Stefanie.Kranz@uk-augsburg.de

Dennis Hartmann
dennis.hartmann@uni-a.de

Bernhard Bauer
bernhard.bauer@uni-a.de

Kerstin Schlögl-Flierl
kerstin.schloegl-flierl@uni-a.de

[1] Institut für Arbeitsmarkt und Berufsforschung, Nuremberg, Germany

[2] University of Augsburg, Augsburg, Germany

[3] University of Augsburg, Augsburg, Germany

[4] University Hospital Augsburg, Augsburg, Germany

[5] Center for Responsible AI Technologies, Munich, Germany

designed to support individualized treatment while continuously learning from clinical feedback to improve its performance over time (cf. Kranz et al. 2023).

Combining doctors' expertise with the analytical power of AI could improve diagnostic accuracy (Tschandl et al. 2020). However, technical sophistication alone is insufficient to fully grasp the implications of AI in healthcare. Questions of patient autonomy, clinical decision-making, and trust cannot be resolved solely on functional or technological grounds. Concepts, such as autonomy and human-centeredness, are not neutral universals but are deeply shaped by cultural assumptions and value systems (cf. Ho and Vuong 2025). Le and Ho (2024) caution that technological innovation—especially in healthcare—should not be automatically equated with progress, particularly when it reconfigures fundamental human relationships and the social fabric of care.

Moreover, it is necessary to question the presumption of AI neutrality. Nguyen and Ho (2025) introduce the concept of the "algorithmic unconscious" to highlight how recommendation algorithms often reproduce normative assumptions without making these visible or debatable to users and developers. What appears as a neutral suggestion may in fact reflect hidden cultural biases or design choices that go unacknowledged. These reflections point to a broader concern: the acceptance and evaluation of AI in medicine is not merely a matter of technical performance or diagnostic precision but also shaped by social mechanisms of idea selection and rejection. Such processes are embedded in collective dynamics of meaning-making and rarely follow purely rational logic. Instead, they are shaped by implicit norms, institutional path dependencies, and the broader cultural economy of knowledge production.

Additionally, the meaningful use of AI in clinical contexts—particularly where trust and responsibility are central—necessitates a participatory approach that actively involves those affected by these technologies. Especially in light of the challenges outlined above, such a participatory framing is not merely normative, but essential for addressing the so-called black-box problem, for fostering trust, and for identifying potential biases in the implementation of AI systems. As highlighted by the German Council of Experts on Health (SVR), AI-based decision support systems may hold particular promise in diagnosing complex constellations of symptoms (SVR 2021, p. 73)—yet their effectiveness depends on how well they are embedded in lived medical practices.

Participation, however, must be understood in a broader sense—not limited to individual consultations but extended to the design, deployment, and governance of these systems. The WHO has stressed the importance of equitable technology transfer and inclusive development practices, particularly in the context of the Global South (WHO 2021, p. IX). In this project, we approach participation not only as a method of inclusion but also as a comparative and reflexive practice: What does it mean to design AI systems that are sensitive to diverse social contexts? In line with recent recommendations by the German Ethics Council on the use of Clinical Decision Support Systems (CDSS), we also draw attention to the notion of system responsibility. Accountability, in this view, must extend beyond the individual doctor to include actors at the meso- and macro-levels—such as regulatory bodies, institutions, and developers. Only when responsibility is collectively anchored can it serve as a meaningful support structure for doctors and patients alike (ZEKO 2023, Recommendation A11).

The OCTOLAB project responds to these complexities through an embedded ethics and social sciences approach (McLennan et al. 2020), led by the Center for Responsible AI Technologies (CReAITech). This approach ensures that ethical and societal implications are considered throughout development (cf. Jörg et al. 2023). As part of this initiative, we conducted scenario-based focus group discussions with citizens of Augsburg to gather qualitative data on their expectations, needs, hopes, and concerns regarding AI-based medical technologies.

This paper asks: *How can citizen perspectives be meaningfully integrated into the design and ethical governance of AI systems in healthcare?* Using BCC care as a case study, we examine how social imaginaries and lived expectations can inform the development of AI systems that are not only technically robust but also socially acceptable and ethically grounded.

A particular focus is placed on the model of shared decision-making (SDM), a framework that foregrounds the collaborative and dialogical nature of clinical decision processes between doctors and patients. SDM seeks to balance clinical expertise with patient values and preferences, thereby operationalizing the ethical principles of autonomy and beneficence within the clinical encounter (Charles et al. 1997; Elwyn et al. 2012). It entails not merely the transmission of information but a bidirectional negotiation of meaning and responsibility, in which both parties contribute to the formulation of an informed, context-sensitive treatment decision (Makoul and Clayman 2006; Joseph-Williams et al. 2014). Within this paradigm, the introduction of AI presents a double-edged transformation: while AI can enhance SDM by providing structured, data-driven recommendations and facilitating evidence synthesis, it simultaneously introduces novel tensions concerning explainability, patient autonomy, and the redistribution of epistemic authority between human and algorithmic agents (Longoni et al. 2019; Riganti and Hoffmann 2024). Understanding how citizens and patients perceive these shifting dynamics of trust, agency, and accountability in AI-supported decision-making thus constitutes a central concern of this contribution.

The following sections first outline the methodological design of the focus groups, and then explore key themes that emerged from the discussions—including the evolving doctor-patient relationship, patient agency, and the specific context of medical practices. We conclude by discussing how these perspectives can inform the development of AI systems in healthcare.

2 Methodology

This study employed a scenario-based focus group methodology to integrate citizens' perspectives into developing AI-based systems for SDM in medicine. One focus group was conducted in Augsburg, Germany, and participants were recruited through local advertising, including posters and flyers distributed throughout the city. A total of 13 individuals participated in the session, which lasted approximately 2 h. This study received ethical clearance from the Ethics Committee of the Ludwig Maximilian University of Munich (LMU Munich) on April 18, 2023 (*Ethikkommission der Medizinischen Fakultät* der LMU München).[1] Participants provided written informed consent, ensuring compliance with ethical standards for research involving human subjects.

The decision to use a scenario-based focus group methodology was driven by the need to understand how citizens perceive the potential use of AI in a medical context. The scenario-based approach enables the exploration of complex issues, such as ethical concerns, privacy considerations, and the perceived impact on patients' autonomy, in a structured yet flexible way. This approach is especially valuable in technology-driven fields like AI, where citizens' views may vary significantly depending on their prior knowledge and the perceived risks and benefits.

We specifically chose the scenario-based approach as outlined by Felt and Fochler (2008), which highlights the importance of incorporating citizens' perspectives in developing technologies. Felt et al. argue that while much attention has been paid to new forms of governance and public participation, relatively little empirical focus has been given to the public's perception of the models, possibilities, and limitations of such involvement and governance. Their work emphasizes the need for a technology-sensitive approach to public engagement, which we consider particularly relevant in the medical domain and seek to explore further. We aimed to explore citizens' views on how AI should be used in medicine, particularly in diagnosing and treating BCC, to ensure that their perspectives inform the development of AI tools that align with their needs and ethical concerns. Additionally, this methodology is effective in engaging participants in discussions that go beyond abstract opinions, allowing them to respond to concrete scenarios where AI is directly applied in medical practice. This approach provides a rich context for understanding how citizens evaluate AI's role in healthcare, which is essential for developing patient-centered, ethically sound AI systems.

The focus group methodology serves as an effective means of integrating diverse members of the public, who can be analyzed as a "mini-public" (Goodin and Dryzek 2006). Through their participation in iterative discussions, this approach allows for the examination of emerging technologies through the lens of public narratives. In this context, the methodology provided valuable insights into public perceptions of AI in medical settings. While our focus group was heterogeneous in terms of gender, age, and educational background, the sample was racially homogeneous, as all participants identified as white. Additionally, the recruitment process was self-selective; participants responded to flyers and posters, suggesting a pre-existing interest in AI. Furthermore, we intentionally did not collect data on disability status due to privacy concerns, making it impossible to assess whether individuals with disabilities were represented. These factors raise important questions about the generalizability of our findings to broader populations, particularly given that AI systems for medical diagnostics are often trained on datasets that predominantly feature white individuals. As a result, the perspectives captured in our focus group may reflect specific cultural and social contexts that do not necessarily represent the broader public's views on AI in healthcare. However, the objective of this study was not to generalize but rather to gain in-depth insights into how a particular group of citizens—who may have varying degrees of familiarity with AI and the healthcare system—engage with the use of AI in diagnosing basal cell carcinoma. More importantly, our aim was to explore *how* these concerns could be integrated into the development of AI systems. Nevertheless, future studies should actively seek to increase diversity among participants, particularly regarding race, disability status, and other intersecting factors, to ensure broader applicability and a more comprehensive understanding of public perspectives.

The three scenarios presented in the focus group were specifically designed to reflect different ways AI could be integrated into the diagnosis and treatment of BCC, each presenting distinct ethical, professional, and societal questions. These scenarios were developed through a collaborative process involving both medical and AI experts, ensuring that the technological aspects were both realistic and applicable to current AI capabilities in dermatology.

The session began with an introduction to AI, encompassing both technical and philosophical perspectives. This introduction aimed to provide participants with a foundational understanding of AI, enhancing the depth and relevance

[1] Project Number: 23-0130.

of the ensuing discussions. Specifically, it covered how AI can be defined from both engineering and philosophy-of-technology standpoints, including distinctions between traditional rule-based systems and data-driven approaches. Key concepts, such as machine learning and deep learning, were explained, with a focus on how these technologies process data, learn patterns, and produce probabilistic outcomes.

In addition, the introduction presented the specific AI use case developed within the research project. Participants were shown visual representations of the prototype combining OCT with a long-pulsed infrared laser for the diagnosis and treatment of BCC, as well as the intended integration of AI for diagnostic support and therapy optimization. We explained how the AI component functions, what kind of data it uses, and what types of output it is meant to offer. This contextualized the discussion and allowed participants to engage with AI not as an abstract concept but as a concrete, medically relevant application.

Following the introduction, three different scenarios were presented and discussed:

- Scenario 1: This scenario was aligned with the goals of the OCTOLAB project, which involves the integration of OCT with a long-pulsed infrared laser for the diagnosis and treatment of BCC, enhanced by AI-assisted diagnosis and therapy optimization.
- Scenario 2: In this scenario, the AI-based diagnostic and therapeutic tool is not used by doctors but by trained nurses. It was intended to explore potential shifts in professional responsibilities within the healthcare system, such as task delegation, accountability, or implications for medical education. However, in the course of the focus group discussions, these issues were addressed only marginally and did not emerge as prominent themes. While some participants briefly mentioned the need for specialist knowledge and appropriate qualifications, the scenario did not prompt an in-depth debate on responsibility or systemic role changes. As a result, this topic did not form a separate category in the grounded theory analysis, which focused on the core concerns articulated by participants in AI-supported medical practice.
- Scenario 3: This scenario assumes that the AI-powered device could be purchased and used by individuals at home, raising questions about accessibility, equity and social class, user autonomy, and the impact on traditional healthcare.

These scenarios were designed to reflect a wide range of potential real-world applications of AI in healthcare—from clinical environments to decentralized, individual use. They were informed not only by conceptual considerations but also by qualitative interviews conducted with doctors involved in the OCTOLAB project. These interviews, held in strict confidence and not analyzed within this paper, provided valuable insights into how practicing doctors themselves imagine future scenarios of AI integration in medical practice. Due to concerns about anonymity and confidentiality, these data are being evaluated in separate publications and are not included in the current analysis.

We acknowledge that the design of these scenarios likely influenced the statements and responses made by participants. However, we believe that this orientation, grounding the scenarios in medically and technically plausible pathways voiced by doctors, offered a productive framework for discussion. Presenting the scenarios sequentially allowed participants to build upon previous reflections, engage with increasingly complex questions, and articulate their perspectives in relation to shifting contexts of AI-supported decision-making. This structure made it possible to surface concerns that are both specific to particular settings and reflective of broader normative questions in human–AI interactions.

While the scenarios were not formally pilot-tested, they were reviewed by domain experts in AI, dermatology, and bioethics, ensuring their relevance and plausibility. Moreover, the focus group itself functioned as a form of scenario validation, allowing participants to critically engage with and assess the plausibility, implications, and desirability of each case.

Each scenario was introduced sequentially, followed by in-depth discussions among participants. The session lasted approximately 2 h, and was recorded, transcribed, and anonymized for subsequent analysis.

For the analysis of the focus group data, we employed Constructivist Grounded Theory (Charmaz 2006). This approach was suitable for our aim of understanding how participants make sense of AI in medical decision-making, as it focuses on how meaning is co-constructed between researchers and participants.

We started by coding each transcript line by line, noting key phrases, recurring ideas, and emotional reactions. In this phase, we identified a wide range of codes, including concerns about trust in AI, unequal access to care, uncertainty about medical roles, and attitudes toward transparency and control. Each transcript was coded independently by two team members, who then met to compare and refine the codes.

In the next phase, we used focused coding to identify patterns across the data. We constantly compared statements from different participants and across groups. For instance, trust was often mentioned in relation to how much personal interaction with doctors participants expected. Comments on agency were more variable and ranged from people

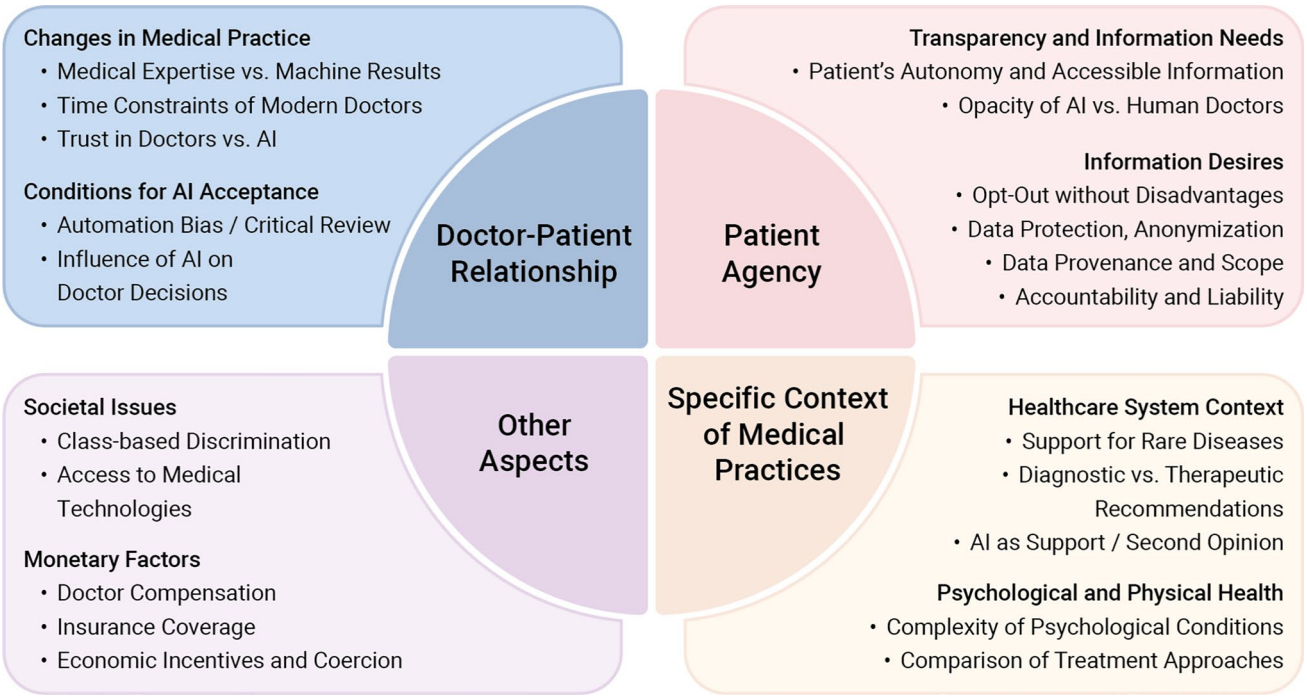

Fig. 1 Overview of key aspects identified in participant discussions

wanting more say in decisions to preferring clear guidance from professionals.[2]

Through this iterative process, three central themes emerged that helped us structure the discussion:

- The **doctor–patient relationship**, especially regarding communication, trust, and acceptance;
- **Patient agency**, including their ideas about the kind of information required to make informed decisions and maintain agency;
- The **specific medical context**, such as differences between physical and mental health or diagnostic and therapeutic use cases.

These three themes were not predefined but emerged through the iterative process of coding and analysis. Other issues—such as data privacy, commercial interests, or the role of regulation—were also discussed by participants and appear in our empirical material. However, we chose to center our analysis on the above three dimensions, because they consistently cut across the discussion and provided the most productive structure for understanding both SDM and the relational, ethical, and contextual concerns that citizens voiced. We thus see these themes not as an exhaustive summary of all concerns raised but as a conceptual scaffolding that best organizes and interprets the empirical richness of the data in relation to our project's aim: supporting SDM in AI-based diagnosis and treatment of BCC.

To increase methodological transparency and illustrate the process of theme development, Table 1 presents selected examples of how participant excerpts were transformed through initial and focused coding into the final thematic structure.

In the next section, we provide a detailed analysis of these citizen perspectives and present our empirical findings in depth.

3 Citizen perspectives on the OCTOLAB project goals

In this section, we present the findings from our discussions with citizens. Given our commitment to prioritizing citizen perspectives in this paper, we adopt a largely descriptive approach in this section. In contrast, Sect. 4 will take a more analytical stance, integrating these perspectives with insights from our interdisciplinary team and existing literature.

The structure of Sect. 4 is designed to correspond directly with the themes explored here: Sect. 4.1 will build upon 3.1, 4.2 upon 3.2, and 4.3 upon 3.3.

[2] A more detailed overview of the subthemes and their interrelations is provided in Fig. 1 (p. 7). This figure illustrates how the broader categories were broken down and how they guided the structure of the next sections.

Table 1 Illustrative examples of the constructed grounded theory coding process

Excerpt (raw data)	Initial code	Focused code	Final theme
‚I can maybe enter into an artificial dialog, but it doesn't replace a real conversation with a doctor.'	AI lacks emotional component; preference for human interaction	Communication and trust in AI contexts	Doctor–patient relationship
‚Nevertheless, there should be basic information, opportunities for further personal learning, the option to refuse the use of AI as an aid in diagnosis…'	Desire for transparency; right to refuse; informed consent	Preserving individual autonomy	Patient agency
‚Diagnosis, definitely not. It's about recommendations. The doctor must still conduct their examination as usual, but now they have a second opinion.'	Different expectations for AI in diagnosis vs. therapy	Role of AI depends on clinical phase	Specific medical context
‚Do doctors have the choice of implementing the AI or not? Is it being introduced by health insurance companies? […] How is the whole thing linked to money?'	Financial incentives; systemic influence	Commercial and institutional pressures	Other aspects

While Sect. 4 primarily focuses on the technical and medical dimensions of AI development—examining how certain citizen concerns may be addressed through design choices, transparency measures, and clinical integration—we are equally aware that not all fears, expectations, or ethical tensions can be resolved through technical means alone. Many of these issues point to broader social, cultural, and institutional challenges that call for political debate, public education, and sustained civic engagement.

In particular, the aspects we have grouped under the category of "other aspects" (Fig. 1) represent concerns that cannot be adequately addressed within the scope of AI design or development processes themselves. Rather, they point to areas where societal transformation and structural change are required. We therefore turn to these dimensions briefly here, before moving on to our main analytical categories in the next subsections.

One of the key themes that emerged during the discussions was the participants' concern with issues of justice and financial inequality, which were particularly pronounced due to the nature of the scenarios presented. The third scenario, where individuals could purchase and use the AI-enabled device at home, raised questions about competency and provoked critical discussions about capitalism. Central to these concerns were fairness and social inequality issues, as participants highlighted the potential for exacerbating class-based discrimination.

> *„I can imagine that, to put it even more pessimistically, it will divide our society. The wealthy can afford the good medicine, and that will probably be done by highly trained doctors in a very transparent, protected, non-abusive area with AI. The others will somehow have to use self-diagnostic tools at home to call the telemedicine.*" (pos. 170, speaker: B2)[3]

The idea that access to such advanced medical technologies could become a privilege reserved for the wealthy was a recurring theme. Participants expressed fears that socio-economic divides could deepen, as only those who could afford the device would benefit from its advantages. These concerns extended beyond individual access to the device itself; they also encompassed broader questions about how AI might reshape healthcare delivery. Participants speculated whether doctors would receive higher compensation for incorporating AI into their practices and how insurance companies might cover these new technologies. Concerns arose about the potential for AI to become mandatory in medical practice, driven by its perceived efficiency and cost-effectiveness rather than patient-centered outcomes. As one

[3] All quotes are originally in German. They have been translated into English by the authors.

participant said, "Do doctors have the choice of implementing the AI or not? Is it being introduced by health insurance companies? Do they get a different reimbursement if they use AI in their diagnosis? How is the whole thing linked to monetary aspects?" (position 84, speaker: B9).

The concerns raised by focus group participants are not merely speculative but closely align with broader scholarly debates on health equity and the commercialization of AI. Wealthier patients are more likely to benefit from increased access, while lower income groups may remain reliant on overstretched public healthcare services (Leslie et al. 2021; Obermeyer et al. 2019). Participants' discussions on whether AI would primarily serve financial rather than medical interests resonate with longstanding critiques of the commercialization of healthcare (Starr 1982). While AI is often framed as a tool to enhance efficiency, persistent concerns remain that its implementation may be guided more by economic incentives than by patient-centered outcomes (Morley et al. 2020). These reflections also played a central role in how participants constructed trust (or mistrust) in AI-supported care: if technologies are seen as embedded in a system driven by profit motives, this can undermine trust not only in the AI itself, but also in the professionals and institutions promoting its use (Herrmann and Pfeiffer 2023). Such concerns thus underscore the complexity of integrating AI into healthcare, which forms a crucial context for the more specific topics explored in the following sections.

Following our analytical framework, this section systematically examines the participants' concerns, needs, expectations, and hopes. We delve into how these perspectives reflect broader societal attitudes toward integrating AI in medical contexts, particularly concerning SDM in diagnosing and treating BCC.

The following figure summarizes the various aspects identified by participants during their discussions: the doctor–patient relationship, patient agency, and the specific context of medical practices, which will be discussed in the subsequent sections.

3.1 The impact of AI on the doctor–patient relationship

Throughout the discussions, participants frequently revisited how the direct medical practice they experience during individual appointments might change if AI were to become an integral part of diagnosing and treating diseases, particularly basal cell carcinoma in this case. A recurring theme was the nostalgic image of the doctor–patient relationship of the past, although it remained unclear which specific period this referred to. One participant expressed it: "In the past, it was said that the family doctor should feel their patient. A machine cannot do this; it operates on a different level. A sick person and a client are concepts that do not come together" (pos. 43, speaker: B4). This sentiment reflected a perceived loss of a more personal, empathetic interaction in modern medicine, and participants contrasted this with the contemporary doctor, often seen as pressed for time: "Is this doctor nice to me, or does he not have time?" (pos. 32, speaker: B4). This observation ties into the broader critique of the tension between seeing patients as human beings versus treating them as customers and the time constraints imposed by profit-driven healthcare systems. The participants feared that even without AI, clinics already struggle to provide sufficient time for patients, and the introduction of AI could exacerbate this issue.

Interestingly, participants frequently compared doctors to machines in their discussions about AI. One participant remarked, "I have no idea what the accuracy rate is when I go to the doctor" (pos. 112, speaker: B5). At the same time, another noted, "Humans can't retain or process more than three or four parameters at a time. Maybe AI can help here" (pos. 32, speaker: B4). This highlights how participants also evaluated doctors using similar rationales, like technical parameters, akin to how machines might be assessed.

However, participants also identified the limitations of AI and recognized the value of human doctors. One participant stated, "But to make decisions, no. There are 20 other factors that a doctor can somehow sense if they are a good doctor" (pos. 182, speaker: B4). This illustrates the participants' belief that there is more to medical decision-making than the knowledge that can be encoded into AI, even though this additional knowledge is difficult to define and was described as something a doctor can "somehow sense." Furthermore, trust was a key factor that participants felt could not be replicated by AI: "There's also a psychological aspect, small things, but when I go to the doctor, do I trust this doctor or not?" (pos. 41, speaker: B4). One participant added, "I can maybe enter into an artificial dialog, but it doesn't replace a real conversation with a doctor" (pos. 117, speaker: B13). This underscores the importance participants placed not only on receiving accurate treatment but also on experiencing trust and the ability to ask questions in a human-to-human interaction.

Participants also engaged in nuanced discussions about the conditions under which a doctor–AI relationship might be acceptable. They examined how this relationship should be structured to gain public acceptance. One participant suggested: "I would say an important point is who sees the patient first. First, the AI-equipped camera, and then the doctor already has some pre-generated information. AI becomes an extension of medical tools" (pos. 42, speaker: B13). We observed here that the order in which the AI system is used in medical practice plays a significant role in the participants' evaluation. This participant raised concerns about the influence of AI on doctors: "If the doctor has the AI diagnosis in front of them, will they be strongly influenced

by it?" (pos. 42, speaker: B13). These comments reflect concerns that doctors might rely too heavily on AI due to time constraints, lack of alternatives, or habit without critically assessing the AI's decisions. In the academic discourse, this risk is often called *automation bias*. One participant articulated this perspective and expanded on it: "But then it also comes down to whether doctors, as mentioned earlier, are under time pressure, whether they blindly trust the AI's decision or recommendation? And there's also the important question of whether doctors fully understand how the AI works. Can they handle it?" (pos. 62, speaker: B8). This comment highlights the importance of how AI systems are *developed* (explainable) and *integrated* into medical practice (as a support system, doctors are not under time pressure). In Sect. 4, we will look closely at this topic and bring together the technical and medical perspectives and requirements.

3.2 Information and agency: citizen perspectives on AI in patient care

Our analysis revealed that participants had varied perspectives on the transparency and information requirements concerning AI in their medical care. While some citizens did not find the so-called "black box" problem of AI troubling, comparing it to their limited insight into the inner workings of human doctors, others emphasized the need for accessible information and autonomy in decision-making.

Several participants equated the opacity of AI with the inscrutability of human doctors. One participant remarked, "So this is a black box. And as you just said, every person is a black box; I don't know how many of you come to your results" (pos. 45, speaker: B5). They added, "So this black box is nothing new" (EBD.). A more detailed comparison was offered: "But whether this doctor, this doctor is competent or makes good diagnoses, I have no idea, and I just trust that since the person has studied, it will be fine. And with AI, I think that for these cancer diagnoses, AIs are currently much better than doctors. We set a bar that is just impossible to reach. I find that interesting" (pos. 113, speaker: B6). These participants discussed whether we extend undue trust to doctors, despite knowing little about them. Trust, as discussed here, is primarily derived from institutional practices and credentials, such as the completion of medical education and the fact that a clinic has employed the doctor. They questioned whether such reliance on institutional validation could lead to overestimating a doctor's competence, particularly in contrast to emerging technologies like AI.

Some participants did not express significant concerns about the 'black-box' nature of AI but emphasized the importance of access to certain types of information and the opportunity to educate themselves about AI. It can be inferred that patients interpret the concept of the black box in contrast to an absolute understanding: if something is not a black box, they should be able to understand and see everything, though this is not necessarily required. However, from their perspective, certain information is essential to make an informed decision and promote patient agency. One participant stated, "From a patient's perspective, it's important to know what happens before this process. That I receive information about the use of AI and also have the opportunity to learn more about it" (Pos. 44, Speaker: B9). Another participant underscored the need for both basic information and opportunities for further personal research: "Nevertheless, there should be basic information, opportunities for further personal learning, the option to refuse the use of AI as an aid in diagnosis, and transparent communication during the consultation with the doctor" (Pos. 50, Speaker: B13). Several participants highlighted the importance of maintaining the option to refuse AI involvement in their care. They stressed that those choosing to opt out of AI-based technologies should not face any disadvantages. "In this scenario, I want to have the choice for it to work without AI. And my concern is that, due to its many advantages, it may no longer be possible to opt out" (Pos. 53, Speaker: B10). The participants expressed a desire not only to reject AI systems in their treatment but also for precise information and transparency. Importantly, their concept of transparency was not focused on technical details but rather on gaining genuine insight into medical practices and ensuring honest communication from doctors and medical staff. They wanted access to information or, at the very least, the opportunity to obtain further details, along with the freedom to make informed decisions about their care.

The *types* of information citizens in our focus group desired were further specified in several instances. Some participants were mainly concerned with the origin and scope of the data used to train AI systems. One participant remarked, "And in this regard, I see it as the right of the patient or the treating doctor to know, first and foremost, what data the AI I am using has been trained on" (Pos. 66, Speaker: B13). These concerns also extended to the potential for bias and discrimination due to underrepresentation in the training data. One participant pointed out, "This is also more difficult for people with darker skin in Germany because they are underrepresented in the training data" (Pos. 71, Speaker: B7). Overall, participants seemed to equate larger datasets with improved outcomes. One participant stated, "When it comes to assessment through imaging and the AI has access to a large dataset, I would trust it more or at least not object to the AI making a recommendation." Thus, their desire for more information regarding the training data was tied to both the representation of subgroups and the size of the dataset.

In addition to these concerns, participants raised specific demands regarding data protection. As one participant emphasized, "Compliance with data protection is essential,

and the anonymization of data, which is often problematic, as medical practices are not always well-trained in data privacy" (Pos. 84, Speaker: B9).

3.3 How the specific medical context of AI application influences citizens' needs

The previous two analytical categories addressed citizens' discussions and statements that are highly relevant to our project and could be broadly applicable to all medical AI contexts. In this third category, we aim to highlight how a specific medical context shapes the needs and concerns of the participants. Although our scenarios and overall project focused on the diagnosis and treatment of BCC, the participants rarely discussed the specifics of BCC diagnosis or treatment. This seemed natural to us, as none of the participants may have had personal experience with BCC and were likely more interested in the general application of AI in medicine. Nevertheless, we identified several aspects where a specific medical context significantly influenced the participants' evaluations.

The broader healthcare context—whether hospitals, for instance, operate as profit-driven institutions—had a notable impact on how AI was perceived, as reflected in the nuanced assessments of the participants, as discussed in Sect. 3. One participant noted: "But also, in which healthcare system is it used? I think that's another big issue. Am I being examined because I have symptoms or pain, or is AI being used to check if someone has something, even without symptoms?" (Pos. 56, Speaker: B2). Other participants similarly addressed the specific context of AI use. There was a clear distinction in their acceptance of AI when it made diagnostic recommendations versus therapeutic recommendations for doctors. Diagnostic recommendations were generally more accepted, while therapeutic recommendations were viewed with greater skepticism: "These devices are good for examinations to suggest a suspicion. But deciding what it is, and even more so for treatment—God forbid" (Pos. 68, Speaker: B4). This distinction was also evident in our second scenario, where trained nurses could use the AI device: "I would differentiate between diagnosis and therapy here" (Pos. 179, Speaker: B5).

Some participants argued that it is essential to differentiate between the application of AI in mental versus physical health conditions, while others disagreed. One participant observed: "With a mental illness, it's so complex, and I don't know if neurological imaging would be used. But I would find it much harder for AI to make an appropriate assessment of a mental illness" (Pos. 35, Speaker: B5). Conversely, another participant stated: "And if both physical and mental health issues are based on large datasets, probabilities, and certain mechanisms, then I'm almost inclined to categorize them the same way. Of course, I'm a fan of having the doctor give the final opinion, but it doesn't make a difference, which is a bit paradoxical" (Pos. 38, Speaker: B10).

Throughout the discussions, participants repeatedly emphasized that AI should be considered a tool to assist doctors rather than an autonomous decision-maker. One participant stated: "It should be an aid for specialists. Nothing more, nothing less" (Pos. 215, Speaker: B4). Another echoed this view: "Diagnosis, definitely not. It's about recommendations. The doctor must still conduct their examination as usual, but now they have a second opinion" (Pos. 216, Speaker: B6).

In conclusion, our analysis revealed that for some participants, the specific context in which AI is applied—such as in mental health—makes a significant difference. In contrast, for others, it is less relevant. However, all participants agreed on two key points: diagnostic recommendations from AI were viewed less critically than therapeutic recommendations. AI should always be used as a support tool, not as an independent decision-making entity without doctor oversight. In the next section, we will integrate these findings with the development of our AI system for the diagnosis and treatment of BCC, incorporating both technical and medical perspectives.

4 Integrating the citizen perspective into AI-assisted shared decision-making: requirements for AI development from a technical and medical perspective

Section 3 discussed three analytical categories that highlight citizens' expectations, hopes, and concerns. We aim to incorporate these insights into developing and integrating our AI system for diagnosing and treating BCC. This section will include a discussion of our team's technical and medical perspectives, following the structure laid out in Sect. 3, beginning with the doctor–patient relationship.

This chapter pursues two main objectives: first, to situate our empirical findings within the existing literature from the fields of medicine and technology; and second, to explore how the concerns raised by our participants can be meaningfully taken into account in the design and implementation of our AI system, thereby addressing our central research question. These two levels—conceptual reflection and practical application—are addressed in each of the following subsections, allowing us to present both critical discussions and specific design decisions.

4.1 Medical and technical perspectives on the doctor–patient relationship

From a medical standpoint, concerns about how the increasing use of AI might affect the doctor–patient relationship are

understandable. The value of a trusting, compassionate relationship between a patient and their doctor is irreplaceable. This relationship is rooted in communication, understanding, and the personal connection from knowing the patient—a cornerstone of medicine for centuries. AI should not replace this connection but rather complement it. The goal is to provide healthcare professionals with more precise diagnostic tools, enabling the earlier and more accurate detection of skin conditions like BCC (Tschandl et al. 2020). The power of AI lies in its ability to analyze vast amounts of data and identify patterns that may not be immediately apparent to the human eye , which could lead to earlier detection of skin cancer and more personalized treatment plans, ultimately improving patient outcomes (Haggenmüller et al. 2021). Therefore, if the AI system is used as a supportive tool, it could potentially strengthen the doctor–patient relationship rather than weaken it. To strengthen these potentials and address the identified needs, the medical team involved in this project has decided to reflect on and enhance the training sessions that are already part of clinical practice. In particular, the risk of automation bias, as highlighted by the participants, will now become an integral component of this training.

There is speculation that AI might allow doctors to focus more on the human aspects of care—addressing patients' concerns, discussing treatment options, and ensuring patients feel supported. AI could alleviate some of the time spent on repetitive tasks, enabling doctors to focus more on the patient as a person rather than just the disease. However, both participants and we are concerned that the time saved by AI could instead be used for economic reasons, such as increasing patient throughput in clinics. The impact of AI on doctors' time is uncertain and remains a critical issue that requires further exploration. Importantly, this promise—that automation would free up time for more meaningful, human-centered work—is not new. Historical studies of earlier waves of workplace automation have shown that such expectations often go unmet. In many cases, the time savings achieved through automation were redirected not toward enhancing quality of care or worker well-being, but toward efficiency and productivity goals (Noble 2011; Zuboff 1988). These patterns raise important questions about whether AI will genuinely enhance clinical care or simply reinforce existing systemic pressures toward acceleration and optimization.

The desire for transparency—frequently articulated by our focus group participants—was also directly linked to the integrity of the doctor–patient relationship. Although participants often expressed limited concern about the AI being a 'black box' in a strictly technical sense, they repeatedly emphasized that doctors should be able to understand and explain how AI systems arrive at their conclusions. This suggests that transparency is valued less as an abstract technical feature and more as a communicative practice that enables trust. The ability of a doctor to explain an AI-based recommendation in understandable terms was seen as central to maintaining SDM and preserving patients' sense of agency. In this way, explainability becomes embedded within the relational fabric of clinical care, rather than merely a property of the algorithm.

In response, our interdisciplinary team worked to translate these concerns into concrete design and implementation goals. The technical subteam explored explainable AI (XAI) techniques, such as feature visualization, saliency mapping, and model attribution methods, aiming to produce outputs that are not only technically interpretable but clinically meaningful. Methods like SHAP (Shapley Additive Explanations), as proposed by Lundberg and Lee (2017), have been used to determine feature importance in tree-based models, as demonstrated by Khater et al. (2023) in a skin image classifier. Other techniques, like gradient-based Class Activation Mapping (CAM) developed by Selvaraju et al. (2017), have been employed to highlight diagnostically relevant areas in skin images, improving the interpretability of convolutional neural networks (CNNs) (Jiang et al. 2021; Mridha et al. 2023). However, our approach moved beyond technical metrics of interpretability to also address how these explanations are used in practice.

Medical professionals in the project provided critical input on what constitutes a 'useful' explanation in the clinical workflow and what kinds of information are needed to support transparent communication with patients. These insights informed not just model outputs but also how AI findings are presented in clinical interfaces. Furthermore, discussions within the team highlighted the risk of automation bias and professional deskilling if explainability is not designed to reinforce, rather than replace, medical judgment. These concerns led to the inclusion of training modules for doctors and to the adoption of hybrid decision-making models that emphasize doctor oversight and responsibility.

This perspective is supported by recent research in dermatology AI, which shows that saliency methods and other XAI techniques can both support and distort clinical reasoning, depending on how they are implemented and interpreted (Winkler et al. 2019; Stieler et al. 2021; Wang et al. 2022). Our project directly addresses these tensions by fostering close collaboration between developers and doctors. Doctors were involved not only in the annotation of training data but also in co-designing explanation formats and providing feedback on their clinical relevance.

With this approach, we sought to implement another key component in our design: a certain level of data literacy, which we have conceptualized elsewhere as *XAI literacy* (Ziethmann et al. 2025). We define XAI literacy as a set of competencies that enable individuals to critically evaluate the outcomes of XAI methods and, in doing so, to exercise a

broader form of AI literacy. At the core of this competence lies an awareness of the limitations of XAI—and, by extension, of AI systems themselves—within specific domains of application, such as dermatology. Each field presents its own constraints on explainability, and it is crucial to recognize that AI systems cannot identify or communicate their own epistemic boundaries.

These limitations become apparent in several key areas:

(1) Inherent data bias—the datasets underpinning XAI outputs are never neutral; all data, and consequently all AI models, reflect particular social and institutional perspectives.
(2) Correlation versus causation—AI systems generate predictions based on patterns rather than causal mechanisms, which can lead to misleading interpretations if not critically assessed.
(3) Discriminatory parameter effects—features that influence AI decisions, such as gender, may have unintended discriminatory consequences if decontextualized. While gender is often treated as a problematic or irrelevant variable in other AI domains, in medicine, it can hold genuine clinical relevance and therefore requires particular attention.

For the purposes of XAI literacy, it is sufficient to acknowledge that no dataset or AI system can ever be entirely neutral, regardless of how neutrality is defined. This recognition establishes the necessary epistemic boundary for practical education in explainable AI. Ultimately, we contend that only doctors who possess XAI literacy will be able to communicate information to patients and citizens in ways that meet the needs and expectations expressed in our focus group.

4.2 AI development to enhance patient agency

In Sect. 3.2, we summarized the key demands voiced by citizens in our focus group to ensure they feel informed and empowered to participate in medical decision-making. Participants emphasized their trust in doctors, expressed the desire for access to information, and stressed the importance of having the option to decline AI-assisted treatments. They also raised specific concerns about data protection and algorithmic bias. These concerns are closely tied to fundamental questions about agency: Who remains in control when AI enters the clinical space? Who understands the system, and who can opt out?

Patient agency—the ability of individuals to actively participate in decisions about their care—is a foundational value in contemporary medicine. It is closely linked to the principles of autonomy, informed consent, and SDM. Research shows that patients who are engaged in decision-making tend to experience higher satisfaction (Shay and Lafata 2014), better adherence (Stacey et al. 2024), and improved outcomes (Krist et al. 2017). However, the introduction of AI into medical contexts can both support and threaten these ideals, depending on how systems are designed, explained, and communicated.

Citizens in our study expressed a desire not just for technical transparency, but for meaningful understanding—particularly when AI is used in diagnostics. They called for information that would allow them to interpret the role of AI in their treatment, rather than merely accept its presence. This suggests that explainability is not only a technical design goal but a communicative practice essential for patient empowerment. When AI is presented as a 'black box', patients are excluded from the interpretive process, which can undermine their sense of participation and control. Conversely, when AI outputs are accompanied by understandable explanations—visual, verbal, or probabilistic—patients are better able to evaluate those outputs, ask questions, and make informed decisions in collaboration with their doctors.

For this reason, our technical medical teams have decided to introduce visualizations, such as saliency maps or heatmaps, to our medical imaging. These highlight the regions that the AI considers to be most relevant, helping both doctors and patients to understand why a specific recommendation is being made. Such tools allow patients to question, confirm, or decline AI-influenced diagnoses, thereby reinforcing their role as decision-makers rather than passive recipients. Likewise, we have provided confidence scores and probability estimates to convey the uncertainty inherent in AI predictions, offering patients an additional interpretive layer that supports agency rather than blind acceptance. As Alkhwaldi (2024) suggests, transparency about the certainty of AI predictions is critical to user trust and uptake.

However, explainability alone does not guarantee agency. As several participants in our study noted, algorithmic bias remains a serious concern—particularly regarding the representation of People of Color in training datasets. These concerns are especially pronounced in fields like dermatology, where the training datasets used to develop AI models may not adequately represent diverse populations, resulting in algorithmic bias (Navarrete-Dechent et al. 2018; Adamson and Smith 2018). Studies have shown that AI systems can perform poorly when diagnosing skin cancer in individuals with darker skin, as these groups tend to be underrepresented in the training data (Aggarwal and Papay 2021; Daneshjou et al. 2022). If patients are not informed about these limitations, their autonomy may be compromised. A system that appears transparent but is silently biased can reinforce existing disparities while offering the illusion of neutrality.

To address this, our interdisciplinary team advocated for layered forms of transparency: not only showing how

a system works, but also clearly communicating its limitations and known biases. For example, interfaces might include warnings such as: '*This model has been trained predominantly on images of lighter skin tones. Diagnostic performance may vary in individuals with darker skin.*' Such disclosures are not merely technical caveats but ethical imperatives. They acknowledge patients' right to understand the potential risks and inequities embedded in the technologies shaping their care.

Bias mitigation was another central theme in our development discussions. Computer scientists emphasized the importance of diversifying training datasets by collecting dermatological images from underrepresented populations and using data augmentation techniques. However, our team also recognized that more diverse data alone cannot eliminate all forms of bias. A model trained on heterogeneous data may still perform unevenly if it is optimized for the majority case—typically lighter skin tones. One proposed solution was to train demographically specific models; each tailored to particular skin types. This approach, though more complex, may better ensure equitable diagnostic accuracy across diverse patient populations.

In parallel, federated learning was discussed as a promising strategy for training AI models across multiple institutions without centralizing sensitive patient data (Haggenmüller et al. 2024). This method not only supports data protection—another citizen priority—but also enables the aggregation of more diverse data sources, potentially enhancing both fairness and model generalizability.

Ultimately, transparency, explainability, and fairness must be understood as conditions for patient agency in an AI-supported clinical environment. Agency requires more than informed consent in a legalistic sense—it demands that patients can understand, question, and meaningfully participate in decisions that involve AI. This means recognizing AI not just as a medical tool, but as a social component that reshapes how information is communicated, how trust is negotiated, and how care is practiced (Gill 2022).

Our project integrated these insights by involving medical and technical experts in iterative discussions that continually returned to concerns voiced by our participants. Doctors reflected on how their roles might shift as they are increasingly tasked with interpreting AI outputs and communicating them to patients. As concrete outcomes, the training sessions were revised to incorporate the concerns and topics raised by the participants—most notably, the issue of automation bias. Developers worked to align model transparency not just with clinical workflows, but also with ethical commitments to inclusion and autonomy. By grounding system design in patient-centered values, we aimed to support an AI integration process that enhances, rather than undermines, the core principles of SDM.

4.3 Context-aware AI development for BCC diagnosis and treatment

As discussed in Sect. 3.3, the acceptability and perceived legitimacy of medical AI are shaped not only by general principles like transparency or fairness, but also by the specific medical context in which AI is applied. In our project, we focused on BCC as a concrete case, and the citizen panels made clear that their expectations of AI vary depending on whether it is used for diagnosis or therapy, and whether it is applied to physical or mental health domains. These distinctions are essential when translating societal concerns into development priorities for AI systems in dermatology.

A particularly striking finding was the clear distinction that many participants made between diagnostic and therapeutic applications of AI. While AI-based diagnostic support was widely accepted—often understood as a 'second opinion'—AI involvement in therapy was met with greater skepticism. This surprised us, given that the system we are developing also includes therapeutic recommendations, such as whether laser therapy is appropriate and how it should be calibrated based on the diagnostic results. We believe that this distinction may be due, in part, to a lack of familiarity: AI-based diagnostic tools are more commonly represented in public discourse, whereas therapeutic applications may have evoked associations with fully automated treatment, potentially without doctor involvement. In several discussions, it became apparent that 'therapy by AI' was perceived not simply as a treatment *suggestion* but as a step toward depersonalized or even autonomous care delivery—something many participants rejected. This highlights a gap in understanding that we, as a project team, have not yet fully addressed. We acknowledge that we need to communicate the role of doctors in interpreting and implementing therapeutic recommendations more clearly and emphasize that our AI system is designed to support, not replace, clinical decision-making. How we integrate this feedback into the development and communication strategy remains an open question currently under discussion in the project.

From a technical standpoint, BCC diagnosis through medical imaging lends itself well to current AI methods, particularly CNNs, which have shown strong performance in classifying skin lesions (Esteva et al. 2017; Brinker et al. 2018; Tschandl et al. 2020; Adegun and Viriri 2021; Haggenmüller et al. 2021; Shetty et al. 2022). Yet, as participants rightly suspected, such performance depends on large, labeled datasets—an area where dermatology still faces limitations (Goyal et al. 2020). In response to the various concerns raised by participants, we have decided to address this issue, particularly in the case of rare diseases and imbalanced data sets, using a promising approach: active learning (Stieler et al. 2023). Active learning differs from supervised learning in that the AI system actively selects

data points for labeling, rather than learning from a static dataset. Unlike reinforcement learning, it does not rely on reward-based feedback but on targeted expert annotation. In the context of a CDSS, this involves an AI system querying the doctor for labels based on a query strategy and learning from the provided feedback. This iterative process allows the AI system to continuously improve while keeping the doctor in the training loop, ensuring that medical expertise directly informs model development—an aspect that some citizens emphasized repeatedly: AI should learn from doctors, not replace them.

Equally relevant to context was the question of who uses the AI and where. Some citizens drew lines between AI used by dermatologists in specialist settings and AI used by non-doctor personnel or in primary care—distinctions that were, of course, influenced by the progressively layered scenarios we presented. However, the fact that participants responded differently to each scenario directly informed our own assessment of the intended implementation settings and the level of clinical expertise required for its responsible use. These insights are highly valuable for the long-term development of the project.

Crucially, these insights remind us that AI development cannot be detached from its context of use. While our system focuses on BCC, the lessons from our citizen engagement go beyond this specific condition. They show that people differentiate between use cases in nuanced ways—diagnosis vs. therapy, mental vs. physical health, expert vs. non-expert users—and that these distinctions must be reflected in development priorities. An AI system that fails to account for such contextual sensitivities may work technically but falter in practice due to a lack of trust, perceived legitimacy, or clinical appropriateness.

5 Conclusion

This contribution has explored how public perspectives can be meaningfully integrated into the ethical and technical development of AI in dermatology, using the OCTOLAB initiative as a case study. By attending to the interplay between clinical routines, computational architecture, and normative concerns, the project highlights the value of interdisciplinary collaboration in aligning AI innovation with societal expectations. Rather than treating ethical reflection as a post hoc addition to technological progress, our findings underscore the importance of embedding ethical engagement throughout the development process—particularly through participatory formats that bring citizens, researchers, and practitioners into sustained dialog.

Participants in our focus groups expressed a clear preference for AI systems that assist, rather than replace, clinical decision-making. Their concerns centered on transparency, intelligibility, and the continued presence of doctors as accountable agents. AI-generated recommendations were welcomed as second opinions but not as final judgments. Such expectations cannot be addressed solely through better algorithms; they call for institutional safeguards, regulatory clarity, and design processes attentive to the lived realities of care.

In response, the OCTOLAB team adopted an embedded ethics and social sciences approach that seeks to align technical development with clinical expertise and public values. AI was framed not as an autonomous authority, but as a tool shaped by iterative feedback from dermatologists and grounded in situational use. Technically, the project engages with known limitations of dermatological AI by exploring strategies such as diversifying training data and employing privacy-preserving models. These efforts aim to improve both fairness and reliability in real-world clinical settings.

What is ultimately at stake is not just diagnostic accuracy, but the broader reconfiguration of medical authority, patient agency, and epistemic trust in a digital healthcare environment. Concerns around bias, automation, and data governance voiced by participants make clear that ethical implementation cannot be deferred to a later stage—it must be built into the architecture of development from the outset. Interdisciplinary cooperation among doctors, ethicists, social scientists, and computer scientists is therefore not optional, but essential to ensuring that AI supports, rather than undermines, the quality and equity of care.

At the same time, several limitations of our study should be acknowledged. First, the number of participants was small, and the total duration of the focus group discussion was limited. As such, the study does not aim to provide statistically representative findings or to capture the full range of possible concerns related to AI-assisted medical decision-making. Rather, it should be understood as an exploratory study, intended to surface and illustrate a variety of concerns and to examine how such concerns can, in general, be integrated into the development of AI systems. Second, all participants in the focus group were white, a result of our local recruitment strategy via flyers in public spaces. While this was not a deliberate exclusion, it highlights the limitations of self-selection and underscores the need for more inclusive and targeted outreach strategies in future participatory projects. Third, and importantly, our participants were citizens rather than patients directly affected by the medical condition under study. While participants frequently articulated their views from a patient's perspective—drawing on prior experiences with healthcare systems—they may not have had direct experience of the condition (basal cell carcinoma). This helps explain why participants rarely engaged with the specific clinical use case of the system and instead focused on broader questions surrounding AI in medicine. We explicitly acknowledge that

there is a meaningful difference between affected patients and non-affected citizens, and that this distinction has implications for the interpretation of our findings. Finally, our central research question—how citizens' concerns can be integrated into AI development—cannot be conclusively answered within the scope of this study. However, we found that our methodological approach, combining focus group discussions with constructivist grounded theory and iterative feedback into an interdisciplinary development team, was highly effective. It led to concrete design decisions, outlined in Chapter 4, and had a lasting impact on how ethical and social considerations were taken up in the technical design. Based on our experience, we recommend this approach to others, while also suggesting larger-scale studies and more condition-specific recruitment strategies in future work.

In summary, the future of AI in medicine will depend not only on technological sophistication, but also on the ability of development teams to recognize and integrate ethical, social, and systemic challenges into their work. Projects like OCTOLAB demonstrate that participatory, human-centered design is not a luxury, but a necessity—one that can help ensure that AI systems are not only technically robust, but also socially responsive and clinically trustworthy, and ultimately usable in practice. Such approaches have the potential to strengthen medical SDM and uphold values, such as care, autonomy, and justice when implemented effectively.

Author contributions This study was a highly interdisciplinary effort, integrating expertise from ethics, social sciences, medicine, and computer science. All authors contributed to multiple aspects of the research and manuscript preparation. P.Z. and K.S.F. provided expertise in ethical and social science perspectives, leading the development of the methodological framework and the analysis of citizen perspectives. S.K. contributed the medical perspective, ensuring that the study's findings were contextualized within clinical practice and patient care. B.B., F.S., and D.H. brought in their expertise in computer science, particularly in Chapter 4, where the integration of citizen perspectives into AI development was discussed. All authors collaboratively wrote and revised the manuscript, ensuring a comprehensive and interdisciplinary approach.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability The data supporting the findings of this study cannot be shared publicly due to ethical and legal considerations, including participant confidentiality and privacy. However, de-identified data may be made available upon reasonable request to the corresponding author, subject to approval by the relevant Ethics Committee.

Declarations

Conflict of interest The authors declare no competing interests.

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

References

Adamson AS, Smith A (2018) Machine learning and health care disparities in dermatology. JAMA Dermatol 154(11):1247. https://doi.org/10.1001/jamadermatol.2018.2348

Adegun A, Viriri S (2021) Deep learning techniques for skin lesion analysis and melanoma cancer detection: a survey of state-of-the-art. Artif Intell Rev 54:811–841. https://doi.org/10.1007/s10462-020-09865-y

Aggarwal P, Papay FA (2022) Artificial intelligence image recognition of melanoma and basal cell carcinoma in racially diverse populations. J Dermatol Treat 33(4):2257–2262. https://doi.org/10.1080/09546634.2021.1944970

Alkhwaldi A (2024) Understanding the acceptance of business intelligence from healthcare professionals' perspective: an empirical study of healthcare organizations. Int J Organ Anal. https://doi.org/10.1108/IJOA-10-2023-4063

Brinker TJ, Hekler A, Utikal JS, Grabe N, Schadendorf D, Klode J, Berking C, Steeb T, Enk AH, Von Kalle C (2018) Skin cancer classification using convolutional neural networks: systematic review. J Med Internet Res 20:e11936. https://doi.org/10.2196/11936

Charles C, Gafni A, Whelan T (1997) Shared decision-making in the medical encounter: what does it mean? (or it takes at least two to tango). Soc Sci Med 44(5):681–692. https://doi.org/10.1016/S0277-9536(96)00221-3

Charmaz K (2006) Constructing grounded theory. A practical guide through qualitative analysis. Sage, 224 Seiten, London

Daneshjou R, Vodrahalli K, Novoa RA, Jenkins M, Liang W, Rotemberg V, Ko J, Swetter SM, Bailey EE, Gevaert O, Mukherjee P, Phung M, Yekrang K, Fong B, Sahasrabudhe R, Allerup JAC, Okata-Karigane U, Zou J, Chiou AS (2022) Disparities in dermatology AI performance on a diverse curated clinical image set. Sci Adv 8(32):eabq6147. https://doi.org/10.1126/sciadv.abq6147

Elwyn G, Frosch D, Thomson R, Joseph-Williams N, Lloyd A, Kinnersley P, Cording E, Tomson D, Dodd C, Rollnick S, Edwards A, Barry M (2012) Shared decision making: a model for clinical practice. J Gen Intern Med 27(10):1361–1367. https://doi.org/10.1007/s11606-012-2077-6

Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S (2017) Dermatologist-level classification of skin cancer with deep neural networks. Nature 542:115–118. https://doi.org/10.1038/nature21056

Felt U, Fochler M (2008) The bottom-up meanings of the concept of public participation in science and technology. Sci Public Policy 35:489–499. https://doi.org/10.3152/030234208X329086

Gill KS (2022) Actionable ethics. AI Soc 37:1–7. https://doi.org/10.1007/s00146-022-01387-1

Goodin R, Dryzek J (2006) Deliberative impacts: the macro-political uptake of mini-publics. Polit Soc 34:219–244. https://doi.org/10.1177/0032329206288152

Goyal M, Knackstedt T, Yan S, Hassanpour S (2020) Artificial intelligence-based image classification methods for diagnosis

of skin cancer: challenges and opportunities. Comput Biol Med 127:104065

Haggenmüller S, Maron RC, Hekler A, Utikal JS, Barata C, Barnhill R, Beltraminelli H, Berking C, Betz-Stablein B, Blum A, Braun SA, Carr R, Combalia M, Fernandez-Figueras MT, Ferrara G, Fraitag S, French LE, Gellrich FF, Ghoreschi K, Goebeler M, Guitera P, Haenssle HA, Haferkamp S, Heinzerling L, Heppt MV, Hilke FJ, Hobelsberger S, Krahl D, Kutzner H, Lallas A, Liopyris K, Llamas-Velasco M, Malvehy J, Meier F, Müller CSL, Navarini AA, Navarrete-Dechent C, Perasole A, Poch G, Podlipnik S, Requena L, Rotemberg VM, Saggini A, Sangueza OP, Santonja C, Schadendorf D, Schilling B, Schlaak M, Schlager JG, Sergon M, Sondermann W, Soyer HP, Starz H, Stolz W, Vale E, Weyers W, Zink A, Krieghoff-Henning E, Kather JN, Von Kalle C, Lipka DB, Fröhling S, Hauschild A, Kittler H, Brinker TJ (2021) Skin cancer classification via convolutional neural networks: systematic review of studies involving human experts. Eur J Cancer 156:202–216. https://doi.org/10.1016/j.ejca.2021.06.049

Haggenmüller S, Schmitt M, Krieghoff-Henning E, Hekler A, Maron RC, Wies C, Utikal JS, Meier F, Hobelsberger S, Gellrich FF, Sergon M, Hauschild A, French LE, Heinzerling L, Schlager JG, Ghoreschi K, Schlaak M, Hilke FJ, Poch G, Korsing S, Berking C, Heppt MV, Erdmann M, Haferkamp S, Drexler K, Schadendorf D, Sondermann W, Goebeler M, Schilling B, Kather JN, Fröhling S, Brinker TJ (2024) Federated learning for decentralized artificial intelligence in melanoma diagnostics. JAMA Dermatol 160:303. https://doi.org/10.1001/jamadermatol.2023.5550

Herrmann T, Pfeiffer S (2023) Keeping the organization in the loop: a socio-technical extension of human-centered artificial intelligence. AI Soc 38:1523–1542. https://doi.org/10.1007/s00146-022-01391-5

Ho MT, Vuong QH (2025) Five premises to understand human–computer interactions as AI is changing the world. AI Soc 40:1161–1162. https://doi.org/10.1007/s00146-024-01913-3

Jiang S, Li H, Jin Z (2021) A visually interpretable deep learning framework for histopathological image-based skin cancer diagnosis. IEEE J Biomed Health Inform 25:1483–1494. https://doi.org/10.1109/JBHI.2021.3052044

Jörg S, Ziethmann P, Breuer S (2023) MedAIcine: a pilot project on the social and ethical aspects of AI in medical imaging. In: Stephanidis C, Antona M, Ntoa S, Salvendy G (eds) HCI international 2023 posters. HCII 2023. Communications in computer and information science, vol 1832. Springer, Cham. https://doi.org/10.1007/978-3-031-35989-7_58

Joseph-Williams N, Edwards A, Elwyn G (2014) Power imbalance prevents shared decision making. BMJ 348:g3178. https://doi.org/10.1136/bmj.g3178

Khater T, Ansari S, Mahmoud S, Hussain A, Tawfik H (2023) Skin cancer classification using explainable artificial intelligence on pre-extracted image features. Intell Syst Appl. https://doi.org/10.1016/j.iswa.2023.200275

Kranz S, Brunnmeier G, Yilmaz P, Thamm J, Schiele S, Müller G, Key C, Welzel J, Schuh S (2023) Optical coherence tomography-guided Nd:YAG laser treatment and follow-up of basal cell carcinoma. Lasers Surg Med 55(3):257–267. https://doi.org/10.1002/lsm.23638

Krist AH, Tong ST, Aycock RA, Longo DR (2017) Engaging patients in decision-making and behavior change to promote prevention. Inf Serv Use 37(2):105–122. https://doi.org/10.3233/ISU-170826

Le NTB, Ho MT (2024) A review of Robots Won't Save Japan: an ethnography of eldercare automation by James Wright. AI Soc 39:3069–3070. https://doi.org/10.1007/s00146-023-01800-3

Leslie D, Mazumder A, Peppin A, Wolters MK, Hagerty A (2021) Does "AI" stand for augmenting inequality in the era of covid-19 healthcare? BMJ 372:304. https://doi.org/10.1136/bmj.n304

Longoni C, Bonezzi A, Morewedge CK (2019) Resistance to medical artificial intelligence. J Consum Res 46(4):629–650. https://doi.org/10.1093/jcr/ucz013

Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. Presented at the 31st conference on neural information processing systems (NIPS 2017), Long Beach, CA, USA

Makoul G, Clayman ML (2006) An integrative model of shared decision making in medical encounters. Patient Educ Couns 60(3):301–312. https://doi.org/10.1016/j.pec.2005.06.010

McLennan S, Fiske A, Celi LA, Müller R, Harder J, Ritt K, Haddadin S, Buyx A (2020) An embedded ethics approach for AI development. Nat Mach Intell 2:488–490. https://doi.org/10.1038/s42256-020-0214-1

Morley J, Floridi L, Kinsey L et al (2020) From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. Sci Eng Ethics 26:2141–2168. https://doi.org/10.1007/s11948-019-00165-5

Mridha K, Uddin MdM, Shin J, Khadka S, Mridha MF (2023) An interpretable skin cancer classification using optimized convolutional neural network for a smart healthcare system. IEEE Access 11:41003–41018. https://doi.org/10.1109/ACCESS.2023.3269694

Navarrete-Dechent C, Dusza SW, Liopyris K, Marghoob AA, Halpern AC, Marchetti MA (2018) Automated dermatological diagnosis: hype or reality? J Investig Dermatol 138(10):2277–2279. https://doi.org/10.1016/j.jid.2018.04.040

Nguyen DH, Ho MT (2025) On the algorithmic unconscious: can we humanize AI with psychoanalytic principles? Subjectivity 32:55–59. https://doi.org/10.1057/s41286-025-00210-8

Noble D (2011) Forces of production: a social history of industrial automation, 1st edn. Routledge, New York. https://doi.org/10.4324/9780203791806

Obermeyer Z, Powers B, Vogeli C, Mullainathan S (2019) Dissecting racial bias in an algorithm used to manage the health of populations. Science 366(6464):447–453. https://doi.org/10.1126/science.aax2342

Riganti P, Hoffmann TC (2024) Shared decision-making and evidence-based medicine series: exploring contemporary challenges and future directions. BMJ Evid Based Med 29:283–284

Sachverständigenrat zur Begutachtung der Entwicklung im Gesundheitswesen (SVR) (2021) Digitalisierung für Gesundheit: Ziele und Nutzen einer vernetzten Versorgungsstruktur. Gutachten 2021. Nomos, Baden-Baden

Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-CAM: visual explanations from deep networks via gradient-based localization. In: 2017 IEEE international conference on computer vision (ICCV), Venice, pp 618–626. https://doi.org/10.1109/ICCV.2017.74

Shay LA, Lafata JE (2014) Understanding patient perceptions of shared decision making. Patient Educ Couns 96(3):295–301. https://doi.org/10.1016/j.pec.2014.07.017

Shetty B, Fernandes R, Rodrigues AP, Chengoden R, Bhattacharya S, Lakshmanna K (2022) Skin lesion classification of dermoscopic images using machine learning and convolutional neural network. Sci Rep 12:18134. https://doi.org/10.1038/s41598-022-22644-9

Stacey D, Lewis KB, Smith M, Carley M, Volk R, Douglas EE, Pacheco-Brousseau L, Finderup J, Gunderson J, Barry MJ, Bennett CL, Bravo P, Steffensen K, Gogovor A, Graham ID, Kelly SE, Légaré F, Sondergaard H, Thomson R, Trenaman L, Trevena L (2024) Decision aids for people facing health treatment or screening decisions. Cochrane Database Syst Rev 1(1):CD001431. https://doi.org/10.1002/14651858.CD001431.pub6

Starr P (1982) The social transformation of American medicine: the rise of a sovereign profession and the making of a vast industry. Basic Books, New York

Stieler F, Rabe F, Bauer B (2021) Towards domain-specific explainable AI: model interpretation of a skin image classifier using a human approach. In: 2021 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW), Nashville, TN, USA, pp 1802–1809. https://doi.org/10.1109/CVPRW53098.2021.00199

Stieler F, Elia M, Weigell B, Bauer B, Kienle P, Roth A, Müllegger G, Nann M, Dopfer S (2023) LIFEDATA—a framework for traceable active learning projects. In: 2023 IEEE 31st international requirements engineering conference workshops (REW), Hannover, Germany, pp 465–474. https://doi.org/10.1109/REW57809.2023.00088

Tschandl P, Rinner C, Apalla Z, Argenziano G, Codella N, Halpern A, Janda M, Lallas A, Longo C, Malvehy J, Paoli J, Puig S, Rosendahl C, Soyer HP, Zalaudek I, Kittler H (2020) Human–computer collaboration for skin cancer recognition. Nat Med 26:1229–1234. https://doi.org/10.1038/s41591-020-0942-0

Wang S, Yin Y, Wang D, Wang Y, Jin Y (2022) Interpretability-based multimodal convolutional neural networks for skin lesion diagnosis. IEEE Trans Cybern 52:12623–12637. https://doi.org/10.1109/TCYB.2021.3069920

Winkler JK, Fink C, Toberer F, Enk A, Deinlein T, Hofmann-Wellenhof R, Thomas L, Lallas A, Blum A, Stolz W, Haenssle HA (2019) Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. JAMA Dermatol 155:1135–1141. https://doi.org/10.1001/jamadermatol.2019.1735

World Health Organization (WHO) (2021) Ethics and governance of artificial intelligence for health: WHO guidance. World Health Organization, Geneva

Zentrale Ethikkommission bei der Bundesärztekammer (ZEKO) (2023) Klinische Entscheidungshilfesysteme (CDSS)—Ethische Aspekte. Stellungnahme der Zentrale Ethikkommission. Bundesärztekammer, Berlin

Ziethmann P, Hummel A, Althammer A, Schlögl-Flierl K, Heller AR, Brunner JO, Bartenschlager C (2025) From embedded ethics to explainable AI: advancing interdisciplinary collaboration in medical AI development (**preprint**). https://doi.org/10.21203/rs.3.rs-7941605/v1

Zuboff S (1988) In the age of the smart machine: the future of work and power. Basic Books, New York

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