What's *really* important? Priorities of trustworthy AI requirements in healthcare: the perspective of clinicians in Sweden

Hanna Schaff[1]✉[0009−0006−1484−3713], Simone Fischer-Hübner[1,2][0000−0002−6938−4466], Ala Sarah Alaqra[1][0000−0002−6509−3792], Leonardo Horn Iwaya[1][0000−0001−9005−0543], and Farzaneh Karegar[1][0000−0003−2823−3837]

[1] Karlstad University, Sweden
✉hanna.schaff@kau.se
[2] Chalmers University of Technology, Sweden

Keywords: Trustworthy AI · Healthcare · Trade-offs

Abstract. The High-Level Expert Group on AI has defined seven requirements on trustworthy AI. Methods to fulfil the requirements may introduce positive or negative relations between them, and the impact may be context-dependent. Using qualitative research methods and a use case (AI chatbot in healthcare) to contextualise the requirements, we explore end users' perceptions and priorities regarding the requirements, and we synthesise a set of relevant relations. Our results can aid in finding suitable trade-off solutions that consider the end-user when designing trustworthy AI healthcare applications.

1 Introduction

Applications of artificial intelligence (AI) are expected to revolutionise healthcare by increasing efficiency, quality, and personalisation of care. However, ethical concerns exist, e.g. related to data privacy and security, bias in AI, transparency, and accountability, and are known barriers to AI implementation [14, 16]. In the European Union (EU), the EU AI Act addresses some of these ethical concerns through legislation [10], drawing on Ethics Guidelines for Trustworthy AI by the High-Level Expert Group on AI (HLEG) setup by the EU Commission [8]. These HLEG Ethics Guidelines define seven key requirements for the development of trustworthy AI systems, including human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity, non-discrimination, and fairness, societal and environmental well-being, and accountability. While some of these trustworthy AI requirements (e.g. transparency and human oversight) can positively support each other, it is also acknowledged that in implementing methods to achieve trustworthy AI, tensions may arise between the requirements. For example, achieving privacy by minimising the amount of data collected could come at the cost of the accuracy of an AI model's

predictions, if it depends on the quantity and quality of its input data. What these relations, tensions and trade-offs may be is a current topic of research, discussed by works such as [13, 9]. Beyond frameworks and technical developments for addressing trade-offs, users' requirements for addressing them remain to be explored. A holistic understanding of clinician's perceptions and priorities of HLEG's trustworthy AI requirements can help to elicit end user requirements and guide user-centric research and development of medical AI applications.

This paper presents an interview-based qualitative study that aims first of all to investigate clinicians' perceptions of the trustworthy AI requirements in the Swedish context for a use case involving a clinician- and patient-facing clinical AI chatbot for remote monitoring of Parkinson's Disease patients, which is currently explored in the EU Horizon Europe project TRUMAN (TRUstworthy and huMAN-cenrtic AI). We focus on interviewing clinicians in Sweden, because Sweden is a digitally advanced country [11, 29] and has a public health sector that actively invests in medical AI tools [1]. Furthermore, Sweden was the first nation with a national data protection law, has a constitutional principle of public access to information [23], has been considered the most gender-balanced country in the world [15], and social well-being is commonly seen as an important value in Sweden. Hence, the trustworthy AI requirements align with common values in Swedish society and should be considered important by the study participants.

Insights on clinician's perceptions will allow us to derive end user requirements for implementing trustworthy AI principles in our future work. From the insights gained, we aim to infer which relations and trade-offs between HLEG's trustworthy AI requirements may be relevant to clinicians for this use case. To this end, our study also aims to explore clinicians' priorities for the HLEG requirements. We expect that our results on clinicians' priorities will aid resolving conflicts by finding suitable trade-off solutions when designing trustworthy AI healthcare applications. Thus, this study addresses the following Research Questions (RQs). **RQ1:** What are Swedish clinicians' perceptions of trustworthy AI requirements as they relate to the use of AI chatbots in healthcare? **RQ2:** What relations, potential tensions and trade-offs between requirements are relevant to clinicians? **RQ3:** How do the clinicians prioritise among the trustworthy AI requirements, and what influences their priorities?

The remainder of this paper is structured as follows: It first provides background regarding the use case and the HLEG requirements in Section 2, and then presents our methodology (Section 3), the study results (Section 4), discussion of related work, limitations and future work in Section 5, and ends with conclusions (Section 6).

2 Background

2.1 Use case: AI Chatbot for Remote Patient Monitoring

Participants of our study were presented with a use case to consider when answering interview questions, to contextualise the trustworthy AI requirements. The

use case involves a care environment comprising a Parkinson's Disease patient, a caregiver providing daily support, and a clinician responsible for treatment and medical support. The patient is remotely monitored by the clinician through a digital platform. The platform includes movement-monitoring devices attached to the patient's body and a phone app for the patient (assisted by the caregiver) to report nutrition, experienced symptoms, and medication intake. The clinician can view this input data through a phone app and on a website dashboard. An AI chatbot is available on both the patient's and the clinician's side of the application. The patient can use it to get help with reporting on activities and to get non-medical self-care advice. On the clinician's side, the AI chatbot sends summaries of patients' current states, and can be prompted for further information or consultation in clinical decision-making. See Appendix for visual description of the use case.

2.2 HLEG's seven requirements for trustworthy AI

The High-Level Expert Group on AI was appointed by the European Commission. In 2019, they presented their Ethics Guidelines for Trustworthy Artificial Intelligence [8]. These guidelines consist of a framework for trustworthy AI, which defines four ethical principles (respect for human autonomy, prevention of harm, fairness, and explicability) rooted in fundamental rights that all AI practitioners should strive to adhere to. To guide the realisation of trustworthy AI, HLEG translated the ethical principles into seven concrete requirements on trustworthy AI, which are in brief summarised as follows:

1. **Human agency and oversight:** Humans should be involved during development, training, deployment and monitoring of the AI system. Human users should make informed decisions about the system, and the system should ultimately empower its users.
2. **Technical Robustness and safety:** The AI system should be accurate, reliable, reproducible and secure.
3. **Privacy and data governance:** Full respect for privacy and data protection. The AI system should have sufficient data governance mechanisms in place, and should ensure legitimate access to data.
4. **Transparency:** The AI system should be transparent, including on capabilities, limitations, and data access. System decisions should be explained. Humans must be made aware that they are interacting with AI.
5. **Diversity, non-discrimination and fairness:** AI systems should be accessible to all users and should not promote unfair biases. Relevant stakeholders should be involved in the AI system development (and its full life cycle).
6. **Societal and environmental well-being:** The AI system should be environmentally friendly, and its societal impact should be considered to ensure that it benefits all human beings.
7. **Accountability:** There must be mechanisms to ensure responsibility and accountability for AI system outcomes.

The EU General Data Protection Regulation (GDPR) incorporates not only principles for enforcing "privacy and data governance" (including data protection) as defined by the HLEG, but also accuracy, security, transparency and fairness and accountability as fundamental principles for data processing, which are listed in its Art 5, and also requirements for human oversight in its Art 22. For the scope of this paper, we will use the narrower meaning when referring privacy and data protection as defined by HLEG, and refer to other data protection-related HLEG principles, including accuracy, security, transparency and fairness, accountability, human oversight, under their specific names.

3 Methods

3.1 Study Design

The study follows a qualitative research approach. Data were collected through semi-structured interviews with 11 volunteering participants, followed by thematic analysis to derive results. A semi-structured interview format was selected since it is suitable for exploring stakeholders' understanding and preferences, and allows for following up on a participant's ideas [18].

The study was guided by an interpretivist paradigm, recognising that the user experience is inherently subjective and that understanding users' perspectives entails exploring how they interpret the systems they are presented with.

The study was reviewed and approved (HNT 2025/860) by an ethics advisor and the data protection officer at Karlstad University.

Protocol The participants were presented with the use case as context for the study.

We divided and thereby extended the original seven HLEG requirements (see Section 2.2) to nine. The requirement *Transparency* was divided into *Transparency of the AI system* and *Explainability of AI suggestions*, and *Technical robustness and safety* was divided into *Reliability of the AI system* and *Accuracy of AI suggestions*. We determined that the original two requirements encompassed different technical aspects that were likely to vary in importance given the healthcare context, and decided to make the distinction to better understand the stakeholders' priorities.

To guide the semi-structured interviews, the interviewer had a set of 18 questions that were asked to every participant, and 15 follow-up and complementing questions were used to prompt participants dependent on previous answers and available time. The questions did not centre on trustworthy AI requirements in isolation; the aim was to prompt a rich, contextual discussion relating to the requirements. This approach came at the cost of not all requirements being equally represented in the resulting data. The requirements *Societal and environmental well-being* and *Accountability* were not explicitly represented in the guiding questions or in common answers to questions. Open-ended questions exploring

perceived challenges attempted to capture thoughts on these two requirements. All other requirements were discussed throughout the interviews.

The final question tasked participants with rating each of our nine trustworthy AI requirements in terms of its importance to them as clinicians. The rating was done per requirement on a Likert scale from one (not important) to nine (most important). Presenting nine options per requirement allowed (but did not require) the participants to rate requirements distinct values. Each requirement was briefly explained in text when they answered this question, and participants could ask interviewers for further clarification.

Data Collection The interviews were conducted between January 23rd and April 16th 2026. The first author conducted every interview. The same co-author conducted most interviews as second interviewer, and two other co-authors conducted one interview each as second interviewer. Interviews were conducted either via Zoom (a video conferencing program) or in person. Each interview lasted between 30 and 90 minutes. The sessions were voice-recorded, and the recordings were transcribed. One interview was conducted in Swedish, and the rest were conducted in English. The Swedish transcription was translated into English to enable analysis by non-Swedish-speaking researchers. A pilot interview (data excluded from these results) was held before the data collection period and prompted edits and refinement of the protocol. All transcriptions were pseudonymised prior to long-term storage.

3.2 Recruitment and Sampling

Participants were recruited through convenience sampling, using personal and professional networks, as well as public channels such as notice boards at a local hospital, and online contact information for Swedish clinicians or hospital departments.

As the study takes an interpretivist stance and uses Clarke and Braun's thematic analysis, measuring data saturation does not align with our approach [6]. We note that the broad sample, encompassing multiple disciplines, has led to diverse perspectives represented in the data. Across the data, we can identify meaningful alignment among several participants, and we determine that the sample size of 11 participants is sufficient to address our research questions.

Participant demographics is summarised in Table 1. P7 had graduated from medical school (including clinical internships) but was not working clinically. The remaining participants worked clinically in the Swedish healthcare system at the time of the interviews.

Participants' experiences with AI chatbots varied. Five participants (P1, P3, P7, P9, P11) mentioned using AI chatbots for summarising texts and as an alternative to search engines. Six participants (P2, P4, P5, P6, P8, P10) state that they have tried AI chatbots but do not use them regularly (except for Google's AI Overview[1]).

[1] https://search.google/ways-to-search/ai-overviews/

ID	Profession/experience	Gender	Age
P1	Nurse in oncology, PhD	Woman	50-65
P2	Radiologist	Woman	50-65
P3	Cardiothoracic surgeon, CEO of a medtech company	Man	50-65
P4	Speciality trainee (*ST-läkare*) in obstetrics and gynaecology	Woman	18-35
P5	Surgeon	Man	50-65
P6	Licensed doctor in emergency medicine	Man	18-35
P7	Medical graduate, PhD student in pre-clinical research	Woman	18-35
P8	Speciality trainee (*ST-läkare*) in anesthesiology and intensive care medicine	Man	18-35
P9	General practitioner and researcher	Woman	50-65
P10	Physician assistant (*Underläkare*)	Man	18-35
P11	General practitioner	Man	50-65

Table 1. Summary of participant demographics

3.3 Data Analysis

We utilised a thematic analysis method [4, 5]. To capture and consider all relevant ideas, aspects and concepts that participants may describe, initial coding was broad and exploratory. The first author was responsible for conducting the coding and deriving a code book. The code book and coding was then discussed with the co-authors and refined. Themes were derived in joint discussions by the first author and one of the co-authors.

4 Results

4.1 Summary of themes

The following sections present the results of our interviews under three main themes and sub-themes that we derived, which are listed in Table 2.

4.2 Clinicians demand accurate, verifiable, and clinically grounded AI suggestions

Participants stress the need for accurate and reliable AI suggestions. Hallucinations are determined to be a system fault impacting *accuracy of AI suggestions* and *reliability of the AI system. Explainability of AI suggestions* is perceived as a mechanism to counteract the effects of hallucinations, allowing human expertise to revise AI reasoning. To further ensure robust human decision-making, *transparency of the AI system* in terms of system capabilities and limitations must be promoted. High-quality data is seen as a key component of an accurate system, where subjective patient-reported data is less reliable than objective test results. The use of relevant input data sources and diverse training data is also identified as promoting *diversity, non-discrimination and fairness* in the AI system.

Main theme	Sub-theme(s)
Clinicians demand accurate, verifiable, and clinically grounded AI suggestions	i Accurate and transparent suggestions are trustworthy ii Hallucinations threaten utility iii System capabilities and limitations: a communication need iv Subjective data is less reliable for clinical assessment v Awareness of challenges for fairness impacting accuracy
Privacy entails confidentiality and more, but advanced privacy risks are not always well understood	i Confidentiality and control: main privacy concern ii Data minimisation and unlinkability: secondary privacy concern iii Data security invokes concern and expectations iv Trust in experts and certifications as privacy assurance v Low awareness of inference capabilities
Human involvement is seen as essential, but comes with challenges	i Need for human verification of patient-reported data ii AI system's effects on human decision-making iii Clinicians remain accountable, for now iv Considering AI and human abilities v Oversight is insufficient as human presence

Table 2. Mapping of themes and sub-themes

Accurate and transparent suggestions are trustworthy Participants expect the AI chatbot to use high-quality, trusted data sources. Examples include local medical guidelines (P6, P7, P8), peer-reviewed scientific sources (P1, P8, P10) and patients' medical data (P7, P11), since these are believed to result in more accurate AI suggestions. Based on prior experience with commercial AI chatbots, participants (P2, P5, P6, P8, P10, P11) express a desire to verify the AI suggestions. The methods of verification vary between participants. Five participants (P1, P2, P6, P9, P11) describe trustworthy AI suggestions as easy to verify as correct using their own knowledge and expertise, without consulting secondary sources. In the context of a healthcare AI chatbot, six participants (P1, P2, P6, P7, P10, P11) stated that they wanted to know what data the AI chatbot bases each suggestion on, to use for verification of the suggestions.

Hallucinations threaten utility When asked about previous experiences with non-trustworthy AI suggestions, seven participants (P1, P4, P7, P8, P9, P10, P11) described hallucinations. It is clear that experiencing hallucinations significantly damages trust. *"Well, I think the first time it gave me made-up references for papers, then I was actually really shocked and I didn't know that was an issue." -P7* Two participants (P7, P9) expressed adopting more cautious behaviours after experiencing hallucinations. Five participants (P1, P7, P8, P9, P11) described having experienced hallucinations and checked the information against other sources.

We infer that hallucinations threaten utility by misinforming users and by adding verification of AI output to clinicians' workload. To mitigate the risk

of misinforming users, two participants (P9, P10) propose that an AI chatbot should provide a disclaimer when it does not have a reliable answer.

System capabilities and limitations: a communication need AI is perceived as difficult to understand, but its capabilities and limitations are important to be aware of and to account for when making a clinical decision. P3, P9, and P11 described AI as difficult to understand, a "black box" that is impossible to look into. However, P2 recalled a conference presentation of artificial neural networks which fostered confidence in the capabilities of an AI-based diagnostic tool. P5 and P6 similarly requested transparency to understand the capabilities and limitations (including biases) of the software. *"I don't need to know exactly how the AI itself works, but I need to be very aware of the limitations and the strengths of it." -P5*

Subjective data is less reliable for clinical assessment One concern for the use of a clinician- and patient-facing AI chatbot is the reliance on subjective input data, where a common perception is that subjective patient-reported data is difficult to assess.

Four participants (P1, P3, P6, P8) described accuracy of AI suggestions as dependent on quality and quantity of training or input data. P2 reported that commercial AI summaries provided insights from online forums rather than scientific sources, which was perceived as less trustworthy. Five participants (P1, P3, P7, P8, P10) noted a risk that patients could input unreliable data. Patients may struggle to articulate symptoms, leading to misinterpretations; low technical literacy may prevent them from using the input interface effectively, or they may maliciously report incorrect data to impact their treatment plan. Two participants (P3, P6) expressed a preference for relying on objective measurements, such as test results or movement-monitoring data.

We interpret two underlying ideas behind these perceptions. The first aligns AI and human capabilities, determining that data that is difficult for a human to assess (e.g. poorly articulated symptoms) is likely difficult for AI to assess. The second assumes that AI is less capable than humans in identifying and assessing subjective parameters. This aligns with the result that some participants (P5, P7, P10) have higher confidence in AI for providing factual information than reasoning. This is supported by P3 and P4 who expressed concern that AI would likely miss cues and non-verbal communication that a human clinician can capture.

Awareness of challenges for fairness impacting accuracy Participants showed strong awareness of potential hindrances to fair assessments for all patients. Given that equality of access and equal treatment for equal needs has for many years been a goal for the Swedish healthcare system [12], the insights appear to originate in experiences of providing equal care in ordinary clinical practice. Challenges are projected onto the AI-based use-case workflow, indicating that AI is not perceived as an automatic aid in providing equal treatment.

P5 and P7 mentioned the potential for bias in the AI chatbot's suggestions due to input or training data. They both identified that an over-representation of research in a certain group of Parkinson's Disease patients could result in AI predictions that are biased towards that group, providing more accurate responses for them than for other groups. P6 and P8 provided examples of culturally biased input data impacting accuracy, which are likely to appear, as many commonly used AI models are trained with data from the USA or from outside the EU. They describe having been provided AI suggestions based on American or British guidelines, which are not always applicable in the Swedish context due to the availability of and local recommendations for treatment options.

Four participants (P6, P7, P10, P11) expect that a more complex disease profile (e.g. a combination of multiple chronic diseases) would be more difficult for an AI system to assess. P7 reasoned about other aspects of a patient's life that complicate a clinician's ability to provide care: *"There are aspects of their other part of their lives, like financial aspects or social that play a big role in their health and in their lives."* Both examples illustrate the perceived alignment between human and AI (in)abilities to provide equal care.

Challenges related to user interaction with the system were also mentioned. Three participants (P1, P2, P7) highlighted challenges related to patients' low-AI or low-tech literacy and low physical or cognitive capacity to use the system independently as risks for a system that relies on high-quality data.

4.3 Privacy entails confidentiality and more, but advanced privacy risks are not always well understood

Perceptions of *privacy and data governance* risks centre on protecting patient data from unauthorised access. This is perceived to be achieved through proper system design that aligns with patient confidentiality rules and provides sufficient data security measures (encompassed by *reliability of the AI system*). Most participants did not make a clear distinction between deterministic software and AI regarding their potential impact on privacy. This indicates that advanced privacy risks that are unique to AI solutions may not be well understood.

Confidentiality and control: main privacy concern Participants expect confidentiality for patients' personal data through access control and local storage solutions. Five participants (P6, P7, P8, P9, P10) stated that clinicians who are directly involved in a patient's care should have access to their data. P6, P7, and P9 mentioned that approved researchers may also access to the data.

Participants' trust in the AI system provider and its infrastructure varied. P7, P10 considered that the provider could have access to the data to improve the system; however, P10 was adamant that such data should be unlinkable to patients. P7 and P10 expressed concern about third parties accessing data for marketing purposes. P3 and P11 recognised the risk of unauthorised access to data in foreign nations. *"Because if it isn't [stored on Swedish servers], then we can't control how it's used. And our regulations doesn't apply." -P11.*

Data minimisation and unlinkability: secondary privacy concern Ensuring that personal data is not linkable to patients and minimising input data (excluding personal identifiers) are perceived as methods to protect patients' identities.

Four participants (P1, P2, P6, P11) mentioned protection of the patient's identity in the AI chatbot use case. P1 described unlinkability as a method, whereas P6 assumed that the clinician should never input any personal identifiers into the chatbot system (data minimisation). P8 and P9 mentioned a similar practice regarding commercial AI tools, explaining that sensitive personal data should not be input into such systems. P3 noted that sensitive data could be stored, given the correct protection: *"The more personal it gets and the more specific it gets on the specific patient, you really have to protect the data."* These disparate expectations indicate a need to clearly instruct end users on the methods used for privacy enhancement and data protection, to establish which behaviours are necessary to adopt.

Data security invokes concern and expectations *"Who can reach the data? What does the chatbot do with all the information that we give it?" -P6*

Data processing and protection relate to both privacy and data governance, and to computer security mechanisms used to restrict illegitimate data access. P2 elaborated on these concerns, stating that patient data, including clinician information about the patient, must be protected in transit. They raised the question of how to securely de-anonymising it towards people who should have access. P11 called for encryption of data when sending it. P9 concluded that the system should have the same data security mechanisms that the electronic medical records have, and referred to electrocardiogram (ECG) equipment, which use secure data transfer to digital medical systems.

Trust in experts and certifications as privacy assurance Participants were asked what assurances for data protection they would like to have. P2 mentioned Swedac (Sweden's national accreditation body), which accredits medical laboratory equipment, as a source of trust. P9 mentioned SKR (The Swedish Association of Local Authorities and Regions) and TLV (The Dental and Pharmaceutical Benefits Agency), which support and coordinate the implementation of medical technologies across Sweden's independent Regions.

Generally, participants were aware that there are requirements for medical equipment used within the public healthcare system. When a tool is made available to them, it has been approved by relevant authorities and assessed for utility, security, privacy, and more. Further, P1, P6, and P7 mentioned the data protection law as an overarching guarantee (implying trust in providers and hospitals to ensure legal compliance). *"I'm quite comfortable with the laws that we have which regulate who can reach what kind of data and who cannot." -P6*

Five participants (P4, P6, P8, P9, P11) stated that they rely on local hospital engineers, procurement engineers, or tool integrators to verify that there are sufficient security mechanisms and legal compliance. P2 and P10 considered

implementation of a system in other departments, hospitals, or clinics to be a trustworthiness indicator.

Low awareness of inference capabilities To gauge the perceived complexity of patient data in the use case AI system, we asked the participants what type of data about the patients they assumed was processed by the system, how they think the data was processed, and for what purposes.

Participants mainly mentioned the categories presented in the use case description: movement monitor data (P3, P4, P6, P7, P9, P10, P11), symptoms (P1, P3, P6, P7, P9, P10, P11), and medication intake (P5, P9, P10, P11). P2, P5, P6, and P8 mentioned information on living and social situations, such as address, family status, social interactions and habits. P2, P4, P5, P8, and P9 mentioned various other data, similar to what is stored in a medical record: social security number, name, age, height and weight, medical history, family medical history, and medical test results. Conversational history between the patient and the AI chatbot was identified as personal data by P6, P7, and P11.

Only two participants (P2, P6) explicitly mentioned tracking or inference to deduce mental or emotional state. Another participant (P8) describes how the AI chatbot "gets to know" the patient and their habits.

Overall, we observed limited knowledge of tracking and profiling risks, which allow AI to deduce the mental or emotional state of a person, or to conduct psychological or behavioural profiling. This indicates a lack of awareness of the potential of AI-enabled breaches of privacy in complex AI systems, with possible inference of sensitive data, as e.g. discussed in [31].

4.4 Human involvement is seen as essential, but comes with challenges

Human agency and oversight through 1) human verification of input data and 2) maintaining human decision-making ensures that clinical decisions remain robust despite faulty input data. The participants reflected on an AI system's impact on their decision-making and *accountability*. *Societal well-being* is ensured by carefully considering where the AI system can be beneficial.

Need for human verification of patient-reported data Due to risks of patients reporting incorrect data (see Section 4.2), human verification of AI output is required before use in clinical decision-making. *"If you were to manipulate the data into making the system believe that your disease is more severe or less severe than it is, it could lead to different treatment options opening up that you shouldn't really have access to." -P8* P1 noted that this risk exists in a human-only care environment as well; however, P8 elaborated on the belief that a person would experience lying to another human as more difficult than lying to an AI system. Further, P8 and P3 speculated on the possibility of future clinical workflows relying more heavily on AI system assessments of the patients, highlighting the risk of incorrect outcomes due to fabrication of patients' health

conditions. P1 stressed the role of human intervention as a way to combat these risks: *"If the patient suddenly reported very strange symptoms, I wanted to check it with the patients. Is it right? Have you put in the right [thing] or is it some misunderstanding from your side in your report of the symptoms?"* Similar ideas permeate our results on explainability and transparency. Participants express a desire for access to input data to verify the AI system's output.

AI system's effects on human decision-making There is wide agreement among participants that humans must be the final decision-makers, as explicitly stated by P2, P3, P4, and P9. Some participants express how human decision-making might be affected by the AI system. P2 and P9 feared that they will "become uncritical" and "that people stop thinking [for] themselves". P2 speculated about the uncertainty of how a clinician would act in the event of disagreement with an AI chatbot, whether they would engage in discussion and allow it to potentially influence their opinion, or if they would disregard it. P2 also uniquely identified the possibility of bias arising from information provided by the AI system, which could influence the clinician's perception of the patient during in-person meetings (thus indirectly influencing clinical decision-making).

Clinicians remain accountable, for now Several participants (P1, P2, P4, P5) emphasised that they would use an AI-driven tool as they would any other tool; it would be used as a complement to their own expertise, without infringing on their role as clinical decision-maker. Additionally, four participants (P2, P4, P5, P9) said they would not feel obligated to (unprompted) describe how they consult with an AI to their patients, since the clinicians are relying on their own expertise and remain the final decision-makers. This implies that accountability remains with the clinician. P1, P6, and P8 explicitly discussed the current state of accountability when using an AI chatbot in healthcare. P6 expressed concern about who is responsible if a clinical decision is made based on an incorrect AI summary. P8 suggested that at the moment, clinicians are accountable for decisions they make based on information from AI chatbots, but that may be up for change in a future where AI could have a more extensive role in healthcare, and where consulting with an AI chatbot might be mandated. P1 reasoned similarly, speculating that patients may feel a greater sense of safety knowing that an AI chatbot is being consulted, since it holds more extensive knowledge than any single clinician.

Considering AI and human abilities Participants perceive benefits with employing AI in healthcare; however, they also question the extent of an AI chatbot's abilities. We infer that it should be carefully considered for which tasks AI automation is suitable. Eight participants mentioned areas where AI automation can outperform humans or significantly save time, including efficient collection, processing, and presentation of information (P1, P3, P5, P8, P10, P11) and the ability to capture tendencies over long time periods (P2, P6).

However, there are abilities and attributes that are inherently human and that AI is not believed to possess. Participants notably discussed the potential of AI to be flexible and suitable for different types of assessments. P2 questioned whether an AI system could make a necessary, individually adapted decision for a patient even if it diverges from routines or guidelines. Similarly, P4 expressed concern that the AI chatbot could be overly sensitive, unnecessarily recommending patients to seek emergency care. P11 had concerns regarding AI assessments of functional symptoms (symptoms for which an physical cause or underlying disease has been ruled out), believing that the AI might focus on identifying a disease rather than treating the symptoms, which should be the primary focus.

Oversight is insufficient as human presence A recurring concern regarding the use case AI chatbot is the risk of depriving patients and clinicians of human interaction in the healthcare environment. Even when human oversight is ensured, there might be human, intimate aspects to in-person communication that are lost. A sense that the AI chatbot system would feel impersonal to clinicians and patients is shared by P1 and P9. P3 and P7 believe that the system would not fulfil the patients' needs. P3 referred to patients' desire for easier access to clinicians in Swedish healthcare. P7 stated that a meeting between a clinician and patients is sometimes about providing human interaction and support rather than treatment.

4.5 Priorities among the requirements

Table 3 summarises the participants' ratings of each trustworthy AI requirement. All requirements were rated as important, with average scores above six on a scale from one to nine. Mean and median values indicate that *reliability of the AI system*, *accuracy of AI suggestions* and *human agency and oversight* are the three most important trustworthy AI requirements to clinicians in the context of using an AI chatbot in healthcare.

Trustworthy AI requirement	**Mean**	**Median**
Reliability of the AI system	7.3	8
Accuracy of AI suggestions	8.4	9
Privacy and data governance	6.9	7
Diversity, non-discrimination and fairness	6.4	7
Explainability of AI suggestions	7	7
Transparency of the AI system	6.6	7
Accountability	6.7	7
Societal and environmental well-being	6.2	7
Human agency and oversight	8.1	9

Table 3. Participants' ratings of the trustworthy AI requirements

P8 reasoned about reliability and accuracy as a prerequisite to adopting a system: *"If [AI systems] aren't reliable and aren't accurate, I wouldn't really want to use them. [...] So all the other questions wouldn't be important if the first ones aren't fulfilled."* This aligns with findings from Section 4.2 regarding the impact of hallucinations on user trust and the need for objective (perceived as reliable) input data. The demand for human oversight permeates Section 4.4. Participants explained the high priority in similar terms. P2 stated that clinicians need to be able to take control, and P4 emphasised the importance of the clinicians' role as the final decision-maker. P9 summarised the idea: *"Wow well, obviously I think this last one is the most important. I think we always have to stay human... [...] We have to keep control—we humans—and AI is there to support us, but we cannot blindly follow it."*

4.6 Relevant relations and tensions between trustworthy AI requirements

Based on the themes, we identify four relations between trustworthy AI requirements that are relevant to the end-users (especially clinicians) in the use case scenario.

Coupling between fairness and accuracy The participants were aware of challenges for fairness, including the risk that biased input or training data could be reflected in AI output (see Section 4.2). P5 and P7 mentioned biases existing in clinical research, along with the expectation that AI will reflect them. This aligns with previous research on AI biases. For example, [3] shows improved accuracy for AI in gender classification when given stereotyped context. If an AI should reflect current knowledge on a subject (e.g. healthcare), it will reflect the biases inherent to the subject, unless methods are employed to counteract it. The relation between accuracy and fairness is context-dependent, as in the example of bias towards American and British medical guidelines (provided by P6 and P8). In this case, accounting for cultural bias (to increase fairness) would also improve the accuracy of AI suggestions in the Swedish context.

Tension between accuracy and privacy Participants identify high-quality patient data as a prerequisite for high-quality AI suggestions (see Section 4.2). However, participants stressed the gravity of processing personal information and call for data protection measures (see Section 4.3). As both of these aspects were highlighted as important by the participants, we infer that the tension between accuracy and privacy is of interest to clinicians. Additionally, we find that there was some awareness of this tension among participants (P2, P6). When asked about data protection, P2 noted the difficulty of providing patient-specific output in a system relying on unlinkable data: *"That's a bit tricky. How should the system work with the individual when it cannot know the individual?"*

Future impact of efficiency and accuracy on human oversight and accountability We elaborate on P8's speculation about the potential, future extension of responsibilities for AI chatbots in healthcare. Assuming AI chatbot systems for patient monitoring would grow more accurate and able to conduct complex reasoning, it seems a possible development to transfer tasks from the clinician to the chatbot. However, as P8 reasoned, such a development might place the clinician secondary to the AI chatbot, diminishing human oversight and shifting accountability from the person to the system. Since professional displacement is a relevant topic in any domain that adopts AI, and concern for which was voiced by participants such as P2, this is a relevant relation to consider.

Tension between transparency towards clinician and patient privacy From transparency needs expressed by participants, we infer that there is tension between clinicians' transparency needs and patients' privacy. An AI system could be capable of profiling or inferring information about a user (e.g. emotional or mental state determined by a patient's tone of writing). Further, a chatbot could be perceived as a safe space for intimate or personal conversations. Several clinicians expressed wanting access to the input data used by the AI system (see Section 4.2. This conflicts with maintaining patient privacy, if clinicians were to have access to patients' chat histories, or if they would be able to prompt the AI chatbot for intimate details or conversation topics beyond what is relevant to treat the patient.

5 Discussion

5.1 Related work

User-centric AI research within the healthcare domain commonly explores stakeholders' expectations and attitudes regarding the implementation of AI in healthcare workflows [28, 33], perceived barriers and facilitators to AI adoption [14, 16], factors affecting users' trust in these applications [2, 27, 30].

As a by-product, these studies identify a number of ethical considerations perceived as relevant to stakeholders. Other works explore stakeholders' perceptions of ethical aspects arising from the implementation of AI in healthcare exploratorily [21, 24, 32], or focusing on specific issues such as data security or privacy [17, 20, 22], accountability, transparency, bias [25, 7], or fairness and diversity [19, 26]. These studies investigate awareness of ethical issues among stakeholders, or explore perceptions of specific ethical aspects. As far as we can tell, no qualitative studies address stakeholders' perceptions and priorities of a structured set of ethical aspects in healthcare AI.

5.2 Limitations

The small sample size limits generalisability of the results and we maintain that the results indicate perceptions and priorities among clinicians, but cannot be used to draw statistical conclusions.

Perceptions regarding some requirements were prompted more directly than others during the interviews. Specifically, *societal and environmental well-being*, *reliability of the AI system*, *accountability* and *transparency* were less prominent. This has resulted in fewer reflections on those topics, although all except the first were reflected upon by some subset of participants. When rating the trustworthy AI requirements, participants were presented with labels explaining the requirements. We noted a few possible misinterpretations of requirements in participants' reasoning, which could affect the legitimacy of those specific ratings.

5.3 Future work

We started to conduct a comparative study investigating potential cultural differences in the perceptions and priorities of the trustworthy AI requirements with clinicians in Germany. In this paper, findings such as high awareness of fairness issues could be attributed to equality being a core societal value in Sweden. Further, Swedish society is largely digitised, which might increase awareness of underlying mechanisms (training, input data, hallucinations) regarding AI among lay persons. First interviews with five clinicians from Germany already gave indications of regional differences, which we will explore further.

6 Conclusion

Participants demonstrate awareness of ethical issues pertaining to AI, especially in terms of how biased or inaccurate input data affects accuracy, and current limitations in AI chatbots such as hallucinations. Explainability and transparency is seen as a way to support human oversight and agency, especially in case of unreliable input data causing unreliable results. As of now, participants consider themselves accountable for clinical decisions, but speculate about a future shift towards AI as decision-maker. *Reliability of the AI system*, *accuracy of the AI suggestions* and *human agency and oversight* are the highest prioritised requirements by the participants, aligning with their awareness of limitations to accuracy and the idea of human oversight as safeguard. We highlight four relations relevant to clinicians in this use case: 1) relation between fairness and accuracy, 2) tension between accuracy and privacy, 3) future impact of efficiency and accuracy on human oversight and accountability, and 4) tension between transparency towards the clinician and patient privacy. The results can inform development of healthcare AI chatbots that account for end users' needs and priorities regarding ethical considerations and trade-offs.

Appendix

Acknowledgments. The work described in this abstract has been conducted within the project TRUMAN. The research leading to these results has received funding from the European Union's Horizon 2020 Research and Innovation Programme, under Grant Agreement no. 101214000. LHI's work is partly funded by the Swedish Knowledge Foundation (KK-stiftelsen), as well as by Region Värmland (Grant: RUN/230445) and the European Regional Development Fund (ERDF) (Grant: 20365177) in connection with the DHINO 2 project, and by Vinnova (Grant: 2018-03025) via the DigitalWell Arena project.

Disclosure of Interests. The authors report that they have no conflicts of interest.

References

1. AI Sweden: Vårdkartan – en kartläggning av svenska regioners ai-initiativ inom vårdsektorn (2025), https://www.ai.se/sv/nyheter/ny-kartlaggning-stor-ojamlikhet-i-regionernas-ai-arbete-inom-varden, accessed: 2026-04-28
2. Asan, O., Bayrak, A.E., Choudhury, A.: Artificial intelligence and human trust in healthcare: Focus on clinicians. Journal of Medical Internet Research **22**(6) (2020). https://doi.org/10.2196/15154, export Date: 02 April 2026; Cited By: 588
3. Barlas, P., Kyriakou, K., Guest, O., Kleanthous, S., Otterbacher, J.: To "see" is to stereotype: Image tagging algorithms, gender recognition, and the accuracy-fairness trade-off. Proc. ACM Hum.-Comput. Interact. **4**(CSCW3), Article 232 (2021). https://doi.org/10.1145/3432931
4. Braun, V., Clarke, V.: Using thematic analysis in psychology. Qualitative Research in Psychology **3**(2), 77–101 (2006). https://doi.org/10.1191/1478088706qp063oa
5. Braun, V., Clarke, V.: Reflecting on reflexive thematic analysis. Qualitative Research in Sport, Exercise and Health **11**(4), 589–597 (2019). https://doi.org/10.1080/2159676X.2019.1628806
6. Braun, V., Clarke, V.: To saturate or not to saturate? questioning data saturation as a useful concept for thematic analysis and sample-size rationales. Qualitative Research in Sport, Exercise and Health **13**(2), 201–216 (2021). https://doi.org/10.1080/2159676X.2019.1704846, doi: 10.1080/2159676X.2019.1704846
7. Choudhury, A., Asan, O.: Impact of accountability, training, and human factors on the use of artificial intelligence in healthcare: Exploring the perceptions of healthcare practitioners in the us. Human Factors in Healthcare **2** (2022). https://doi.org/10.1016/j.hfh.2022.100021, export Date: 08 July 2026; Cited By: 68

8. Commission, E., Directorate-General for Communications Networks, C., Technology, on Artificial Intelligence, H.L.E.G.: Ethics guidelines for trustworthy AI. Publications Office (2019). https://doi.org/doi/10.2759/346720
9. Duddu, V., Szyller, S., Asokan, N.: Sok: Unintended interactions among machine learning defenses and risks. In: 2024 IEEE Symposium on Security and Privacy (SP). pp. 2996–3014. IEEE (2024)
10. 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) (Jul 2024), http://data.europa.eu/eli/reg/2024/1689/oj
11. Eurostat: Digitalisation in europe – 2025 edition, https://ec.europa.eu/eurostat/web/interactive-publications/digitalisation-2025, accessed: 2026-04-28
12. Fredriksson, M.: Universal health coverage and equal access in sweden: a century-long perspective on macro-level policy. International Journal for Equity in Health **23**(1), 111 (2024)
13. Gittens, A., Yener, B., Yung, M.: An adversarial perspective on accuracy, robustness, fairness, and privacy: Multilateral-tradeoffs in trustworthy ml. IEEE Access **10**, 120850–120865 (2022). https://doi.org/10.1109/ACCESS.2022.3218715
14. Hassan, M., Kushniruk, A., Borycki, E.: Barriers to and facilitators of artificial intelligence adoption in health care: Scoping review. JMIR Hum Factors **11**, e48633 (2024). https://doi.org/10.2196/48633
15. Hofstede, G.: Culture and organizations. International Studies of Management & Organization **10**(4), 15–41 (1980). https://doi.org/10.1080/00208825.1980.11656300
16. Khumalo, A., Brodén, K., Bellström, P., Alaqra, A.: AI adoption in oncology: Implementation process, barriers, and facilitators. In: Selected Papers of the 47th Information Systems Research Seminar in Scandinavia. vol. 15 (2024)
17. Kwesi, J., Cao, J., Manchanda, R., Emami-Naeini, P.: Exploring user security and privacy attitudes and concerns toward the use of general-purpose llm chatbots for mental health. In: Proceedings of the 34th USENIX Security Symposium. pp. 6007–6024 (2025), export Date: 02 April 2026; Cited By: 1
18. Lazar, J., Feng, J.H., Hochheiser, H.: Chapter 8 - interviews and focus groups. In: Research Methods in Human Computer Interaction (Second Edition), pp. 187–228. Morgan Kaufmann, Boston (2017). https://doi.org/https://doi.org/10.1016/B978-0-12-805390-4.00008-X
19. Lee, M.K., Rich, K.: Who is included in human perceptions of ai?: Trust and perceived fairness around healthcare ai and cultural mistrust. In: Conference on Human Factors in Computing Systems - Proceedings (2021). https://doi.org/10.1145/3411764.3445570, export Date: 08 July 2026; Cited By: 123
20. Liu, Z., Hu, L., Zhou, T., Tang, Y., Cai, Z.: Prevalence overshadows concerns? understanding chinese users' privacy awareness and expectations towards llm-based healthcare consultation. In: 2025 IEEE Symposium on Security and Privacy (SP). pp. 2716–2734 (2025). https://doi.org/10.1109/SP61157.2025.00092
21. Meadi, M.R., Sillekens, T., Metselaar, S., van Balkom, A., Bernstein, J., Batelaan, N.: Exploring the ethical challenges of conversational ai in mental health care: Scoping review. JMIR Mental Health **12** (2025). https://doi.org/10.2196/60432, export Date: 08 July 2026; Cited By: 110

22. Mikkelsen, J.G., Sørensen, N.L., Merrild, C.H., Jensen, M.B., Thomsen, J.L.: Patient perspectives on data sharing regarding implementing and using artificial intelligence in general practice – a qualitative study. BMC Health Services Research **23**(1) (2023). https://doi.org/10.1186/s12913-023-09324-8, export Date: 02 April 2026; Cited By: 22
23. Ministry of Justice, Government Offices of Sweden: Public access to information and secrecy – the legislation in brief (2020), https://www.government.se/contentassets/2ca7601373824c8395fc1f38516e6e03/public-access-to-information-and-secrecy.pdf, accessed: 2026-04-28
24. Mirzaei, T., Amini, L., Esmaeilzadeh, P.: Clinician voices on ethics of llm integration in healthcare: a thematic analysis of ethical concerns and implications. BMC Medical Informatics and Decision Making **24**(1), 250 (2024), https://link.springer.com/content/pdf/10.1186/s12911-024-02656-3.pdf
25. Nouis, S.C.E., Uren, V., Jariwala, S.: Evaluating accountability, transparency, and bias in ai-assisted healthcare decision- making: a qualitative study of healthcare professionals' perspectives in the uk. BMC Medical Ethics **26**(1) (2025). https://doi.org/10.1186/s12910-025-01243-z, export Date: 08 July 2026; Cited By: 67
26. Robertson, C., Woods, A., Bergstrand, K., Findley, J., Balser, C., Slepian, M.J.: Diverse patients' attitudes towards artificial intelligence (ai) in diagnosis. PLOS Digital Health **2**(5) (2023). https://doi.org/10.1371/journal.pdig.0000237, export Date: 08 July 2026; Cited By: 111
27. Schlicker, N., Lechner, F., Wehrle, K., Greulich, B., Hirsch, M.C., Langer, M.: Trustworthy enough? examining trustworthiness assessments of large language model-based medical agents. TECHNOLOGY, MIND, AND BEHAVIOR **6**(2) (2025). https://doi.org/10.1037/tmb0000164
28. Solmonovich, R.L., Kouba, I., Lee, J.Y., Demertzis, K., Blitz, M.J.: Physician awareness of, interest in, and current use of artificial intelligence large language model-based virtual assistants. PLOS ONE **20**(5) (2025). https://doi.org/10.1371/journal.pone.0320749
29. The Bertelsmann Stiftung: Digital health index, https://www.bertelsmann-stiftung.de/en/our-projects/the-digital-patient/projektthemen/smarthealthsystems, accessed: 2026-04-28
30. Tun, H.M., Rahman, H.A., Naing, L., Malik, O.A.: Trust in artificial intelligence–based clinical decision support systems among health care workers: Systematic review. Journal of Medical Internet Research **27** (2025). https://doi.org/10.2196/69678, export Date: 08 July 2026; Cited By: 82
31. Vekaria, Y., Canino, A.L., Levitsky, J., Ciechonski, A., Callejo, P., Mandalari, A.M., Shafiq, Z.: Big help or big brother? auditing tracking, profiling, and personalization in generative {AI} assistants. In: 34th USENIX Security Symposium (USENIX Security 25). pp. 8115–8134 (2025)
32. Witkowski, K., Okhai, R., Neely, S.R.: Public perceptions of artificial intelligence in healthcare: ethical concerns and opportunities for patient-centered care. BMC Medical Ethics **25**(1) (2024). https://doi.org/10.1186/s12910-024-01066-4, export Date: 08 July 2026; Cited By: 97
33. Zhang, Z., Genc, Y., Xing, A., Wang, D., Fan, X., Citardi, D.: Lay individuals' perceptions of artificial intelligence (ai)-empowered healthcare systems. In: Proceedings of the Association for Information Science and Technology. vol. 57 (s2020). https://doi.org/10.1002/pra2.326, export Date: 02 April 2026; Cited By: 32