BEHAVIOUR & INFORMATION TECHNOLOGY
https://doi.org/10.1080/0144929X.2024.2334852

Clinical needs and preferences for AI-based explanations in clinical simulation training

Naja Kathrine Kollerup[a], Stine S. Johansen[b], Martin Grønnebæk Tolsgaard[c], Mikkel Lønborg Friis[d], Mikael B. Skov[a] and Niels van Berkel [a]

[a]Department of Computer Science, Aalborg University, Aalborg, Denmark; [b]Australian Cobotics Centre, Queensland University of Technology, Brisbane, Australia; [c]Copenhagen Academy for Medical Education and Simulation, Rigshospitalet, Copenhagen, Denmark; [d]NordSim, Aalborg University Hospital, Aalborg, Denmark

ABSTRACT

Medical training is a key element in maintaining and improving today's healthcare standards. Given the nature of medical work, students must master not only theory but also develop their hands-on abilities and skills in clinical practice. Medical simulators play an increasing role in supporting the active learning of these students due to their ability to present a large variety of tasks allowing students to train and experiment indefinitely without causing any patient harm. While the criticality of explainable AI systems has been extensively discussed in the literature, the medical training context presents unique user needs for explanations. In this paper, we explore the potential gap of current limitations within simulation-based training, and the role Artificial Intelligence (AI) holds in supporting the needs of medical students in training. Through contextual inquiries and interviews with clinicians in training ($N = 9$) and subsequent validation with medical experts ($N = 4$), we obtain an understanding of the shortcomings in current simulation-based training and offer recommendations for future AI-driven training. Our results stress the need for continuous and actionable feedback that resembles the interaction between clinical supervisor and resident in real-world training scenarios while adjusting training material to the residents' skills and prior performance.

ARTICLE HISTORY
Received 9 May 2023
Accepted 15 March 2024

KEYWORDS
Explanations; feedback; learning; clinical; simulation; AI support systems

1. Introduction

Amidst growing global healthcare demands, the training of competent clinicians is critical. Even with advancements in medical procedures, diagnostic errors still present a significant challenge, accounting for up to 17% of adverse events and 10% of cases leading to patient death (Bordini, Stephany, and Kliegman 2017). These errors continue to remain a 'blind spot' in healthcare that rarely are detected. With the continuous improvements in diagnosis and patient care in the health care field, it frequently occurs that diagnostic errors result from human rather than systems errors (Bordini, Stephany, and Kliegman 2017). Considering this, the use of artificial intelligence (AI)-based support systems has been pointed to as a way to overcome these errors. However, AI-support implementations have not yet been prevalent in clinical simulation-based training (Juanes-Mendez et al. 2021). We define AI support systems as systems which 'provide clinicians with knowledge to enhance medical decision-making, such as support for diagnosing patients, making prognostic predictions, or selecting treatments' (Cai, Winter et al. 2019).

Prior medical literature points to increased training and reflection on existing work practices significantly reducing diagnostic errors (Croskerry and Nimmo 2011). Medical training, however, is expensive due to the time required by senior medical staff and hampered by medical reality, in which the challenging and rare cases are handled by the most experienced clinicians with the more repetitive tasks handed to trainees – limiting their practice opportunity (Galvin and Buys 2012). In addition, inadequate training can result in reduced productivity once faced with real-world challenges. Given the vital importance of medical training and the implementation of AI support in healthcare to overcome medical errors (Juanes-Mendez et al. 2021), we explore clinical end-users' identification of limitations in current medical simulation-based training and how AI-driven support can be an integrated part of future clinical simulation-based training.

The need to provide adequate medical training under both financial and time-restricted circumstances has led to using surgical simulators to bridge the gap between theory and medical practice. Through these simulators,

CONTACT Naja Kathrine Kollerup nkka@cs.aau.dk Aalborg University, Selma Lagerlöfs Vej 300, Aalborg 9220, Denmark

students can engage in realistic healthcare scenarios without direct supervision or the risk of endangering patients. Simulators allow users to interactively engage with various procedures using representational tools and instruments similar to those used in examination rooms. To be an effective training tool, surgical simulators must provide effective and actionable feedback to trainees, often residents obtaining necessary clinical skills. Moreover, the surgical simulators must be able to present their outcome in human-understandable terms. Especially in the healthcare domain, the ability of clinicians to understand any suggestion they might act upon is vital for ethical and legal accountability reasons (Panigutti et al. 2022). In addition, interacting with medical simulators without direct supervision provides the need for explanations to guide residents through correct procedures and clinical decisions for future real-life cases. The use of AI for assessing performance, for example, in detecting inefficient or potentially harmful movements by the trainee, has been presented as a means to increase the realism and effectiveness of simulators by more closely resembling a supervisor (Mirchi et al. 2020; Salvetti et al. 2021).

Prior work on emergency medicine training has shown the importance of active learning rather than didactic lectures to 'contextualise content, explain difficult concepts, and improve student learning from simply remembering to applying and analysing' (Wolff et al. 2015, p. 1). It is, however, unclear to what degree current simulators align with the learning needs of medical students and what potential the use of AI-support systems holds for future training purposes. This lack of understanding of end-user needs (both medical students and their clinical educators) with AI feedback and explanations undermines the development of human-centred AI systems as well as the eventual adoption by the intended user group, as seen in a range of malfunctioning medical system (Musen, Middleton, and Greenes 2014; Yang, Steinfeld, and Zimmerman 2019).

Various aspects affect users' usage of AI systems, including trust (Ma et al. 2023), transparency (Wang et al. 2021), and safety precautions (Schneiderman 2022). In this paper, we specifically focus on simulation feedback optimisation through a user-centred design approach to understand how AI feedback can act as a central role in optimising medical simulation training. Given the complex nature of AI systems, we build upon Human-Centered AI (HCAI), in which the AI systems should amplify, empower, and enhance human performance (Schneiderman 2022). Given these considerations, instead of seeing users and computers as a singular entity, we separate the two and see users as the driving force, given that the technology should support rather than constrain users in their everyday work (Preece, Rogers, and Sharp 2015). In close collaboration with two specialised medical training and simulation centres, NordSim at Aalborg University Hospital and CAMES at Rigshospitalet Copenhagen, both in Denmark, we sought to evaluate the role of a simulation-based system to minimise the gap between the current limitations of simulation-based training to future AI implications of training feedback.

To investigate this, we conducted a contextual inquiry involving residents (doctors in training who have a degree in medicine and are now specialising) from gynaecology and general practice. We observed residents interacting with a gynaecology ultrasound simulator as they completed a set of training exercises, followed directly by a semi-structured interview focused on the feedback received throughout the training ($N = 9$). In addition, we engaged with medical experts and presented them with the results obtained throughout our contextual inquiry and interviews to validate and assess the data outcome ($N = 4$).

Our results highlight that the current feedback in the medical simulation diminishes the medical residents' learning practice, given the lack of guidance through feedback. The feedback should be pervasive in the beginning, guiding the residents as a supervisor and slowly subtracting as the medical residents' expertise increases during the practice time. These results provide opportunities and insights into merging the HCAI perspective into the medical simulation training, given that the medical residents stressed the importance of feedback resembling real-life practice and emphasising the support of in-the-moment correction of errors and support. Based on the results, we provide recommendations that emphasise, from medical residents' viewpoint, how AI support can be integrated as a part of medical simulation-based training.

Our work provides an understanding of clinicians' expectations and needs for future implementations of AI explanations in the context of simulation-based training. These insights are critical to constructing human-centred AI systems and building on the growing body of literature on Human-Computer Interaction (HCI) and AI that seeks to ensure that AI feedback is understandable, actionable, and of value to the target audience.

2. Related work

In this work, we emphasise related research within AI support in health care, designing for human-AI collaboration and learning in health care.

2.1. AI support in health care

Due to the increasing technological advancement and availability of AI, various areas in healthcare have sought ways to incorporate AI to support decision making (Cai, Reif et al. 2019; Yang, Steinfeld, and Zimmerman 2019), including patient diagnosis (Cai, Winter et al. 2019), treatment (Yang, Steinfeld, and Zimmerman 2019), and prognosis (Bellio et al. 2021). This integration of AI support systems in the medical domain raises several critical research questions, including trust towards the AI system (Cai, Winter et al. 2019) as well as the interpretability of the system when faced with critical medical decisions (Cai, Reif et al. 2019). HCI plays a crucial role in integrating AI systems into clinical practice by ensuring alignment between user needs, system capabilities, and the presentation of information. Recent work has stressed the importance of incorporating AI technologies into established work practices, highlighting how technically sound systems often fail in aligning with end-user needs (Cai, Winter et al. 2019; van Berkel et al. 2021; Yang, Steinfeld, and Zimmerman 2019).

User-centred design has become prominent in building systems that align with end-user needs. For example, in a study on the onboarding needs of medical practitioners, Cai et al. follow a user-centred approach to identify the types of information pathologists require when first introduced to an AI system (Cai, Winter et al. 2019). Their results show that medical experts desire transparency in the model by creating upfront information about its strengths, limitations, and how the AI was intended to benefit them. Failing to take end-user considerations, such as integration failure, lack of concern for existing workflows, and the collaborative aspect of clinical work, into account during the deployment of technologies will typically result in the non-use of the system. In particular, clinical decision-support systems (CDSS) have suffered from unsuccessful integration in medical practice due to the lack of user-centred design practice (Yang, Steinfeld, and Zimmerman 2019). Rather than deploying a DST system that is disconnected from practitioners' decision-making, Yang et al. introduce the concept of 'unremarkable AI', which challenges the convention of a disconnected walk-up and use systems, instead proposing that AI is situated more naturally in the decision-making process. The significance of this work is embedded in the shift of clinicians' viewpoint towards accepting and trusting clinical DST (Yang, Steinfeld, and Zimmerman 2019).

Given the high-stake decisions that are regularly made in the medical domain, any successful introduction of AI support heavily relies on establishing trust with the end user. When working with AI as a part of the decision-making process in healthcare, the need for successful integration is key for leveraging trust since system failure can cause users to lose trust in the system and abandon it for their domain expertise (Cai, Reif et al. 2019). The absence of trust, for example, as caused by lack of transparency, can severely affect the adaption of AI technologies (Došilović, Brčić, and Hlupić 2018). Interpretability and explainability are two core terms raised when considering AI systems. Mohseni et al. present a multidisciplinary survey on explainable AI (XAI) and how its explanations of AI decision-making process could benefit users (Mohseni, Zarei, and Ragan 2021). The authors share a categorisation of XAI design goals evaluation methods and present a framework where the authors pair design guidelines with evaluation methods (Mohseni, Zarei, and Ragan 2021). Their findings show that the overall design goals for XAI to novice users are to describe how the system works, provide explanations to improve trust, mitigate bias, and raise privacy awareness. Prior research has discussed transparency's benefits in AI systems by providing users with a better understanding of a system's behaviour (Springer and Whittaker 2020). Springer and Whittaker discuss the drawbacks of transparency in AI systems by stating that transparency can result in a loss of confidence in the system, with users trusting it less, leading them to question the system (Springer and Whittaker 2020). In a survey on explainable AI, Došilović defines interpretability as follows: 'Interpret means to explain or to present in understandable terms. Then, interpretability in the context of ML systems is the ability to explain or present in understandable terms to humans.' (Došilović, Brčić, and Hlupić 2018). In a study on AI support for pathologists, Cai et al. demonstrate how introducing an interactive refinement tool increased end-user acceptance and trust compared to a traditional interface (Cai, Reif et al. 2019). In effect, the refinement tool allowed the medical practitioners to understand better how the AI system's recommendation came to be and influence the system's recommendations based on their expertise.

The aforementioned works stress the importance of explainable AI in daily medical practice. In this paper, we focus on exploring the gap between the existing feedback outcomes of simulation-based training environments and how AI-supported feedback can become an essential part of medical training in the future.

2.2. Designing for human-AI collaboration

Building on a rich history of designing collaborative technology in HCI and Computer-Supported Cooperative

Work (CSCW) (Fitzpatrick and Ellingsen 2013; Grudin 1988), recent work shows an increasing interest towards enabling effective Human-AI collaboration. Collaboration is a process of activities typically done between two or more individuals which involve co-management, shared goal setting, understanding, and progress tracking (Wang et al. 2020). Various studies have investigated these different aspects of collaboration, intending to assess how AI systems can help users complete tasks and reach their goals (Oh et al. 2018; Yang et al. 2020). This includes work on the experience of co-creation between an AI agent and users (Oh et al. 2018), human perceptions of AI teammates (Zhang et al. 2021), and reflections on the challenges when designing for Human-AI interaction (Yang et al. 2020). Oh et al. focused on the communicative aspect of AI systems and studied whether providing detailed instructions to users is beneficial in collaborative AI (Oh et al. 2018). Similarly, Schaekermann et al. investigated whether communicating uncertainty can impact cognition and trust (Schaekermann et al. 2020). In their study, Schaekermann et al. explore how AI assistants for medical reasoning should communicate uncertainty about classifying ambiguous cases. Their findings showed that communicating uncertainty can help experts allocate cognitive resources and be helpful during interactions to reassess trust for each case. The efficiency of human-AI collaboration often depends on how humans calibrate their trust towards AI systems. There is an emphasis on the importance of system transparency to maintain users' trust in the AI system (Okamura and Yamada 2020). For example, McComack et al. show how a collaborative AI drummer communicates confidence levels through emotion-based visualisation. This study showcases how transparency, as expressed through visualised cues, can improve musicians' ability to achieve control and engage in successful collaboration (McCormack et al. 2019).

Prior work within Human-AI collaboration has, among other areas, focused on the interplay between transparency and trust, the visualisation of uncertainty to support collaboration, and turn-taking during collaborations (Oh et al. 2018; Okamura and Yamada 2020; Schaekermann et al. 2020; Yang et al. 2020). Highly relevant to our work is prior research on the incorporation of AI systems in learning environments to support learning (Amir and Gal 2013; Mohan, Venkatakrishnan, and Hartzler 2020). Integrating technology, particularly collaborative AI systems, in the learning environments introduces challenges since the systems need to support users in their learning process while minimising the amount of intervention (Amir and Gal 2013). Amir and Gal investigate 'Exploratory Learning Environments', specifically how to visualise students' progress in a chemistry course (Amir and Gal 2013). Teachers use the visualisations generated by their system to assess students' progress and support students in guiding their problem-solving to maximise their learning experiences without interruptions. The authors present an algorithm for detecting students' plans during an exercise and then visualise both this plan as well as the actual temporal course of actions the student took. A similar study conducted by Mohan et al. proposes '[…] a human-aware AI system that collaborates with humans to support their learning and explores how such systems can be evaluated' (Mohan, Venkatakrishnan, and Hartzler 2020). The authors design a coaching agent with two aspects in mind: personal adaptation and temporal adaptation. The authors show that it is possible to develop a formulation for adaptive goal setting for exercises aligned with clinical experts' recommendations.

In this work, we set out to assess the needs of residents in medical training scenarios to understand the required feedback to accommodate learning in the healthcare context. In addition, the role of AI-supported feedback in optimising the medical residents' learning experience will be assessed. Realising the extensive part of collaboration in human-led (medical) training, we build on existing work on designing for human-AI collaboration in identifying the unique requirements imposed in a learning context.

2.3. Learning in health care

Medical education provides a framework of combined theory and practice, training students to become well-rounded medical professionals, after which they typically specialise in a specific area of medicine (e.g. heart surgeon, general practitioner). Simulators play an increasingly prominent role in training practical and procedural skills for individual learning and training to operate effectively in a group setting (Salas et al. 2005). Not only do simulators allow clinicians to practice without potentially harming patients, but they also allow for more control over the learning process (Gordon et al. 2001) – for example, by introducing new material and challenges at a rate that is appropriate to the learner's development (Kneebone et al. 2004). Using simulators in medical training further provides the opportunity to study alternative ways of learning that would be highly unethical when conducted on patients. For example, Dyre et al. instructed trainees in simulation-based training to commit errors instead of the more common instruction to avoid error (Dyre et al. 2017). Their results showed that trainees who purposefully committed errors during training had a higher transfer of learning to the real-world clinical setting.

Likewise, some medical procedures cannot be practised in the clinical setting because of their low frequency – for example, managing cardiac arrest is best done in an environment that allows practitioners to conduct errors and adjust performance without the extreme time pressure that characterises clinical cardiac arrests.

Concerning AI support and explainability, the needs of clinicians may be starkly different in a learning context compared to a real-life clinical context. For example, while AI systems are typically designed to support clinicians in their task (e.g. pathological assessment (Cai, Reif et al. 2019) or polyp detection (van Berkel et al. 2021)), learning often improves only slightly if the correct solution is presented to the learners concurrently during the training. The experience of challenges often benefits the learning process (R. Bjork 1994). Bjork introduces the concept of 'desirable difficulties' (R. Bjork 1994), later described as 'Desirable difficulties […] are desirable because they trigger encoding and retrieval processes that support learning, comprehension, and remembering' (E. L. Bjork and Bjork 2011). Existing clinical support tools aim to avoid introducing such (desirable) difficulties. In other words, what works for improving performance is sometimes detrimental to learning.

In this study, we engage with both the intended end-users of medical simulators (i.e. medical doctors in training – referred to as residents) as well as experienced doctors with a vested interest in medical student training to understand the needs concerning feedback and explanations in clinical simulation training.

3. Method

Through a user-centred research approach, we aim to explore medical residents' needs and expectations for feedback optimisations within medical simulation training to understand how AI feedback can act as a central role in optimising medical simulation training. We have taken a contextual inquiry approach through close collaboration with our medical partners. The contextual inquiry approach consists of observing and interviewing medical residents in their context to gain an understanding of practices and behaviours (Raven and Flanders 1996). In the case of this study, the inquiry acted as a base for a clearly defined set of concerns regarding the exploration of simulation-based medical training, understanding the current feedback provided in the medical training simulation, and the role AI holds for future optimisation in simulation-based training. Our data collection consisted of in-depth semi-structured interviews. In addition, we observed the residents in their interaction with a transvaginal ultrasound simulation prior to each interview, informing our semi-structured interviews. Specifically, we informed a subset of the interview questions based on the challenges observed during the interactions.

In addition, we consulted with medical experts to validate our insights and gain an understanding of the required performance of simulation-based feedback in resemblance to real-life teaching. We, therefore, divided our exploratory study into two stages: exploration and validation. All interviews with residents and medical experts were audio recorded using audio recording software on the primary author's smartphone and laptop. The majority of the interviews were conducted physically face-to-face, while the remainder were conducted over the phone.

Given the user-centred perspective of this study, we do not engage with or evaluate an AI support system. However, we accumulate insights from clinicians to inform future incorporation of AI support systems for clinical simulation training. Participation in this study is entirely voluntary and based on the participants' consent to participate.

3.1. Exploration stage

In the first stage, we explored the medical training environment by observing medical residents' interaction with a transvaginal ultrasound simulation, see Figure 1. In particular, we focused on the feedback provided

Figure 1. A resident interacting with the transvaginal ultrasound simulator.

during training, for example, in the case of errors made throughout the tasks. Observations varied in duration, ranging from one and a half hours to two hours. In addition, we conducted semi-structured interviews following the resident's completion of the training session, in which we focused on the participants' perspectives on the feedback provided by the simulator. This interview aimed to gain more insights into bridging the gap between the challenges of the current use of transvaginal ultrasound simulation and the future incorporation of AI feedback within medical simulation-based training. The interview consisted of four segments: 'understanding the environment', 'training tools in action', 'optimisation in learning', and 'interaction and feedback'. In addition, throughout our observation, we noted time stamps of specific problems the residents encountered during their training. We then observed how the residents navigated these problems and whether the simulation assisted them in overcoming any of the issues encountered. This allowed us to use these observed problems in the semi-structured interview to elicit the residents' interpretation of how they overcame the challenges and if they were aware of their mistakes throughout their training. Our intended goal with the semi-structured interview was not to evaluate the residents' performance during the interaction with the simulation. However, our interest lay in observing how the residents overcame specific problems and how the residents perceived the provided feedback. Given this, the performance reports generated by the simulation were not used as part of our data collection.

We include our interview guide in Appendix 1. At *NordSim* we both observed and interviewed the residents ($N = 7$), whereas at CAMES we were only able to interview the residents ($N = 2$).

The interview guide consisted of an exploratory assessment to understand the training environment, obtain insights into the residents' experience and use of the transvaginal ultrasound simulator, and capture residents' opinions towards possible improvements of the feedback provided to optimise their learning process.

Nine participants from different medical and specialisation areas participated in this stage, see Table 1 for an overview. These residents work towards different specialisation areas, such as general practice and gynaecology. The experience level varied among the residents; some were new to the transvaginal ultrasound procedure, whereas others had limited experience due to real-world clinical work. Each interview lasted for approximately 30 minutes. The exploration stage allowed us to focus on the feedback provided during the interactions and obtain a perspective on how the training contributed to their day-to-day skills as a practitioner.

Table 1. Overview of participant sample from the exploration stage (i.e. first stage of data collection).

ID	Position
P1.1	Introductory position in gynaecology
P1.2	Medical graduate pursuing specialisation in general practice
P1.3	Medical graduate pursuing specialisation in gynaecology
P1.4	Medical graduate pursuing specialisation in gynaecology and obstetrics
P1.5	Medical graduate pursuing specialisation in general practice
P1.6	Medical graduate pursuing specialisation in general practice
P1.7	Medical graduate pursuing specialisation in general medicine
P1.8	Medical graduate pursuing specialisation in general practice
P1.9	Medical graduate pursuing specialisation in gynaecology

3.2. Validation with experts

In the validation stage, we consulted with medical experts to obtain detailed clinical insights on future directions for incorporating AI feedback to support medical residents in simulation training. We presented the experts with initial insights gathered from the contextual inquiry and interviews with the residents and asked them to validate and assess the data outcome.

We conducted semi-structured interviews with four experts, each lasting approximately 30 minutes. Two out of the four interviews with the experts were conducted over the phone, while the other two were conducted in person. We divided the interviews into 'initial understanding' and 'validation of findings'. In the first part, we aimed to obtain an understanding of the experts' position, teaching experience, and work experience), while the second and main part of the interview focused on the validation of our initial findings from our exploration phase. The second part of the interview consisted of four areas: Feedback in the ultrasound simulation, Repetition of tasks, Difference in experience levels, and Rigid assessment of task completion. The lead interviewer read the initial findings under each area aloud for the experts to provide them with a clear understanding of the issues the medical residents raised, followed by a set of questions related to the given area. These questions revolved around the medical experts' perspectives and opinions towards the severity of our insights, e.g. 'What is your view on these ways of learning (see-one-do-one-teach-one and self-directed learning), and how does the focus on giving feedback during practice affect teaching and learning?' Furthermore, we later asked the experts to compare the insights to real-life training and the benefits or detriments of incorporating our initial findings in future simulation-based training. We include our interview guide in Appendix 2.

We are not looking directly at the procedure conducted during the simulation training but at how the simulation can be optimised by providing feedback

Table 2. Overview of expert sample from the validation stage (i.e. second stage of data collection). Years of experience counted from graduating from medical school.

ID	Position	Experience level
P2.1	Chief physician in orthopaedic surgery (UK: consultant)	16+ years
P2.2	Senior physician	10+ years
P2.3	Chief physician in orthopaedic surgery and PhD	16+ years
P2.4	Chief physician responsible for education	18+ years

throughout residents' learning process. Learning and improving your work is specific and centred within all practices of medicine. Therefore, the scope of medical experts was not narrowed to Obstetrician and Gynaecologist experts. We provide an overview of the position and experience of these experts in Table 2.

3.3. Description of the environment

The surgical simulation-based training is located within a simulation centre, where residents can practice their skills within different specialisation areas. The simulation-based training is a mandatory part of their training, as it provides residents with hands-on experience over several hours of practice. The residents who participated in this study have practised in one of two different simulation centres with slightly different approaches to simulation-based training. In each training environment, the medical residents trained and interacted with two identical transvaginal ultrasound simulators. At CAMES, the simulation centre provided residents with additional supervision by an expert through an introduction course ($N = 2$) – this was not the case at *NordSim* . However, a student worker attends the residents' training sessions and assists with any technical challenges regarding the simulation ($N = 7$). This, however, is not considered supervision since the student worker does not have the qualifications to help the residents with any challenges related to the medical practice.

3.4. Transvaginal ultrasound simulator

This study focused on simulation-based training using a transvaginal ultrasound simulator. The primary focus of the ultrasound simulator is to strengthen residents' hand-eye coordination while providing them with a structured approach to performing ultrasounds. The simulator consists of a screen visualising an ultrasound scan of a patient's uterus and a probe to insert into the vaginal canal, resembling real-life practice. An additional screen provides a visual overview of the anatomy and enables residents to track the probe's orientation as they move it, as shown in Figure 2. The probe can be navigated through a simulated vaginal canal while the scan process is displayed as an ultrasound image on one of the two screens. The ultrasound images showcased on the screen are recordings of former patients to provide a realistic case and visuals. Both training environments use the same transvaginal ultrasound simulator model, namely the ScanTrainer Transvaginal Simulator by the UK-based Intelligent Ultrasound.[1]

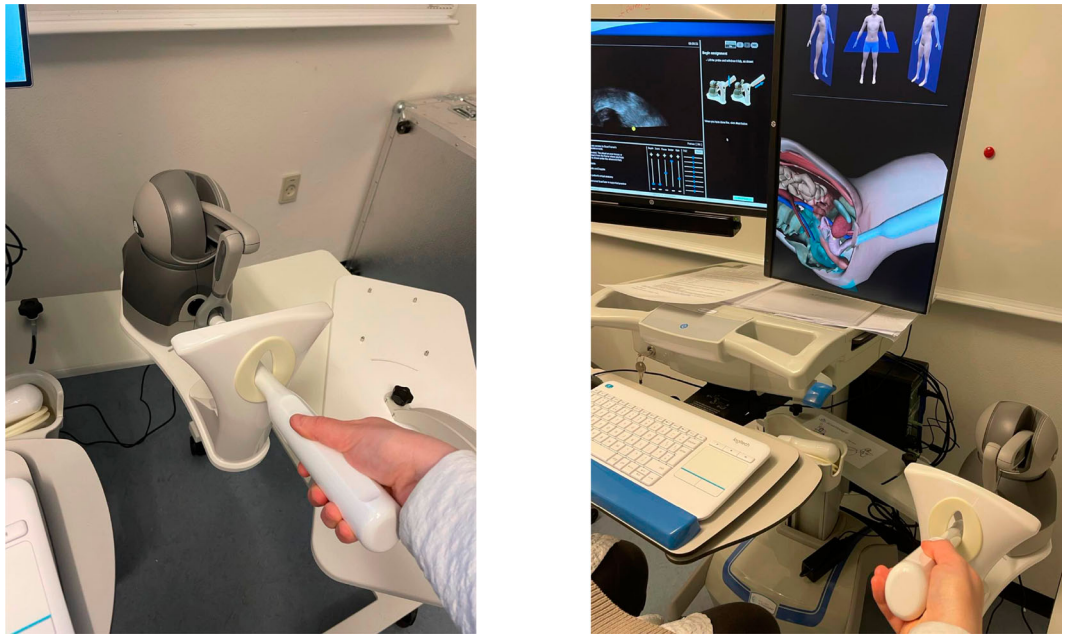

Figure 2. Close-up of the probe and overview of the transvaginal ultrasound simulation.

The simulator provides residents with a task description at the onset of each task and a show-and-tell video visualising the task completion. This video consists of a voice-based explanation of how to perform the task while simultaneously visualising the completion of tasks through a video showcasing the anatomy. Following this video, residents start the task. Feedback is provided to the residents only after completing a training scenario. Here, residents are shown a list of completed and failed tasks and – in the case of failed tasks – a short description of why the resident could not complete the task successfully. If the resident places their cursor on top of a failed task, a picture of the intended outcome is shown together with a description of how to complete the task. Whenever the resident incorrectly handles the probe, the simulator simulates patient discomfort and pain through visual and audio feedback. The indication of pain is connected to a scale with three different pain levels: green (no experience of pain), yellow (mild experience of pain), and red (a high level of pain). Whenever the residents navigate the probe, a cursor will move between the three levels depending on the residents' navigation of the probe – if the cursor moves into the red category, the audio system will produce an 'ouch' sound.

3.5. Data analysis

To avoid disrupting participants during their training session and attain ecological validity as best as possible (van Berkel et al. 2020), no audio or video was recorded during training observation. The data corpus consists of hand-written field notes, photographic images of medical practitioners training with the simulators, and audio recordings from the interviews. All participant quotes subsequently presented in this paper are paraphrased from informal conversations during the interaction or transcribed from the interview data.

We conducted an inductive thematic analysis to identify patterns of meaning across the data set. We follow a bottom-up approach to derive codes and themes from the content of the data, in which there was no attempt to fit the data into an existing theory (Braun and Clarke 2022). As such, through the bottom-up approach, we were driven by the data rather than bringing in concepts to interpret the data – as seen in a more deductive approach to thematic analysis.

Through data familiarisation, we identified codes that captured the residents' perspectives towards the ultrasound simulator's current feedback and insights on their optimisation in health care learning. These codes were initially clustered together into potential broader patterns of meaning related to: feedback; experience; and simulation. These broader patterns contained subpatterns, which initiated the meaning of directions of the three broader patterns. We identified a theme around feedback that appeared to reflect complementary yet distinctively different ideas supporting AI-driven feedback in simulation-based training. We settled on an analytic structure consisting of four themes, where each theme captured the implications for understanding the medical residents' perspective on learning and feedback optimisation in simulation training: use of the simulator; the current state of simulation-based training; and supporting resident development. Four researchers discussed and agreed on the themes, which will be presented in the subsequent section.

4. Results

We frame our findings around three themes we constructed through our thematic analysis: use of the simulator, current state of simulation training, and supporting resident development. Given the expected effect of the studied environment on our results, we first describe the environment in which our observations took place. Our analysis aimed to identify future recommendations for appropriate AI-supported feedback in medical training simulators based on input from residents and medical experts. These recommendations are presented in the subsequent section.

4.1. Use of the simulator

The residents presented several motivations for using the simulator, often highlighting a combination of motivators. The primary motivation is using medical training simulators to strengthen residents' gynaecologic competencies and form a basis for clinical knowledge. The majority of the residents further stressed that the use of simulations allowed them to avoid harming patients while simultaneously providing the opportunity to obtain a rapid systematic approach in conducting ultrasounds;

> Well, you are allowed to stand and try yourself and fumble with it, without there being a patient who lies and has to be a part of the process. Uh, so you have plenty of time to try the techniques and get hands-on with them. I especially think the spacious [referring to the different planes within a patient's abdomen (e.g. coronal plane)] is difficult, so it is just nice to have time for it. (P1.2).

The majority of the residents furthermore distinguish between real-life training (i.e. training in the clinic) and medical simulation training. Here, they emphasised that the basis for learning is knowing their weaknesses in the clinic; '[…] if one does not have a foundation to say that I know my problem, then it will all probably be a little

difficult.' (P1.3). Medical experts, who introduce residents to different procedures, form the basis for learning in clinics. Due to time limitations, however, residents do not get the time required to deal with cases and, therefore, miss out on a lot of training. This specific issue was mentioned by three of the residents, who highlighted that medical experts typically take control during procedures, which causes residents to lack the training needed to accomplish their future tasks accurately and confidently.

> There are some who are bad at taking control and guiding your hand. They [medical experts] put their hands on one's hand because now we must have the picture. You must not do that in my world. You must say it in words and try to guide. (P1.4).

In contrast, residents control the procedure in interacting with the medical simulators. The residents positively commented on the ability to freely submerge in a large amount of training, dealing with atypical cases, which prepares them for real patients in the clinic, and the practice of hand/eye coordination – a beneficial quality within gynaecology.

4.2. Current state of simulation training

The structured approach incorporated in the simulator, in which tasks and cases are predefined chronologically, was generally experienced as positive – with residents highlighting that this resembled the structure of procedures encountered in real-life practice. The predefined tasks allowed the residents to build on a foundation of existing knowledge for further practice and knowledge sharing; 'There's more structure in it; also, you can build more on the foundation in the clinic. Then you have more relevant questions, rather than be confused because it is the second time you have tried it.' (P1.1). The live audio feedback provided by the simulator given indications of pain was criticised by one participant, who mentioned that the simulator's indication of pain is not equivalent to real-life patient care, seemingly indicating that the patients seen in the clinics are more pain tolerant.

> […] because many things are not equivalent to real life, especially the indication of pain, which sets off very easily. But it [the probe] has to go much further in the vaginal canal in reality, and it is not that limited (P1.6).

4.2.1. Feedback

A common issue the residents face is a lack of explanation and instructions during the training session to improve or recover from errors. The limited feedback provided by the system did not provide the residents with an explanation for handling challenges and errors or allowing them to 'ask for help.' This resulted in the residents guessing, completely stopping with the scenario or the simulation altogether, or deliberately making errors to be able to proceed. Here, the residents' primary approach to handling challenges faced during training was to make deliberate errors as they were unable to uncover the correct procedure or, as residents stressed; as deemed to be correct by the simulator. One participant highlighted some of the frustration experienced when faced with unclear instructions: 'I just had to move on. At one point, I had to say, "My best shot is this".' (P1.7). This lack of guidance during their training resulted in voids of crucial information that the residents did not process. Despite this frustration, we observed five out of nine residents engaged with their failed tasks by repeating the entire section of tasks to confirm their learned insights. Residents' reflections often highlight a strong motivation to learn from their mistakes; 'No, there was nothing else but to try again. If it [the task] was not passed, I must try again. See if we can determine what the program wants us to do.' (P1.7). Even though we interviewed residents' from two different simulation-based training centres with varying approaches to training (CAMES with an expert supervision intro-course and NordSim with no additional supervision), surprisingly, we did not observe any difference in the residents' satisfaction related to the need for continuous assessment during training. With the expectation of seeing a difference of opinions from the residents' training at CAMES , they had the exact needs and requirements for feedback to optimise their learning.

Some residents distinguished between 'right' and 'wrong' simulator feedback to determine whether they should repeat a task or move on to the next task. This assessment of the simulator's feedback was based on the residents' expectations towards what was considered acceptable in clinics – measurements that are two millimetres off will not affect patient care and are therefore perceived as 'wrong' simulator feedback. While the simulator's feedback is highly binary (correct or incorrect), residents repeatedly made categorisations as to the severity of their mistakes. For example, a wrongfully taken ultrasound picture was considered major, whereas the wrong placement of a measurement indicator was considered minor. These categorisations influenced the subsequent steps taken by the residents;

> If I could see it had something to do with the picture I had taken, then I would like to go back and check it out. If it had something to do with you not putting the mouse in the right place, then I did not care. (P1.4).

In addition, the presented feedback lacked guidance on the right approach beyond merely indicating mistakes and information on measurement errors. This feedback was often experienced as insufficient by the

residents; 'Well, you get elaborations, in the sense where it says you are four millimetres from where you had to measure, but it does not point to where you should have measured.' (P1.1). The simulator's assessment approach did not meet the residents' expectations since it lacked customised feedback through continuous dialogue and individual evaluation. '[…] in the final examination, I got a minus because I had not correctly centred the picture. So, you could say I needed some personal feedback.' (P1.7) When asked about their criteria for determining the degree of failed tasks, it resulted in conflicting poles regarding criteria for succession within the completion of tasks. Some residents emphasised the severity of the simulator's rigid assessment as unimportant to successful completion. '[…] also, I could see in the results that it was a millimetre I had placed calibres incorrectly. Never mind.' (P1.4). Other residents criticised the lack of transparency in the assessment, as it led to a narrow scope of creativity in the residents' completion of tasks;

> Either you had done what it [the simulator] wanted, or you had not. I think the feedback is like if you take a multiple-choice test. There is a right answer, and there is a wrong answer. There are not so many degrees to it, […] and where one gets the most out of it, I could imagine, if, for example, one had a professional scanner in the room. (P1.7).

4.2.2. Explainability and trust

While the feedback received extensive critique, other aspects of the simulator highlighted residents' trust in the simulator. In particular, the correct detection of patient anatomy was experienced as positive by the residents; 'It is obvious, that it [simulator] can detect the anatomy and everything' (P1.1). Further, the tasks provided by the simulator closely resemble the challenges encountered in the clinic; 'It is nice to encounter the same mistakes. I also think it indicates that it [simulator] is true to reality' (P1.3). One participant stated that to trust the medical training simulator, they expect that feedback is presented correctly and that the simulator is intelligent enough to determine when something is wrong. Even though the residents trusted the simulator, the explainability of the simulation was lacking, with some of the residents having to experiment during the training on how the simulator's algorithm assessed performance. '[…] it took some time before I saw the transparency of the algorithm and how it measured and weighed me. Once I understood it, it was quite easy to pass the tasks. It was not intuitive how it evaluated one.' (P1.7). Moreover, some residents highlighted the need for the simulator to explain and provide criteria for successfully completing tasks, as the simulator only provided the residents with an assessment without highlighting the decision-making process. This lack of explainability resulted in residents perceiving the provided feedback as too rigid and inflexible, with minor mistakes leading to failed tasks; 'So it [the simulator] is very rigid. Just that I had a millimetre difference, I failed.' (P1.3).

4.3. Supporting resident development

The simulator lacked explainability to assess the correctness of the residents' actions throughout the training sessions by not providing a continuous assessment of residents' mishandling of specific tasks, e.g. correcting the way of holding the probe while looking for organs. This builds visible frustration among the residents.

> I needed some personal feedback along the way, where the feedback could indicate how you can change how you hold the probe to find what you are looking for. Sometimes, I was just completely lost in terms of where I was. (P1.2)

During the interviews, most residents highlighted the importance of continuous assessment and advice during practice to assist them in their learning process, prevent errors, and provide reassurance for the residents. As stated by one resident: '[…] I especially think in something like learning or operational, it is important to have ongoing feedback, so you do not get into some bad habits or inappropriate ways of doing it.'.

Despite the majority of residents highlighting the importance of continuous assessments during training, some of our interviews reveal disagreement among the residents as to whether or not a more continuous type of feedback could resemble the learning environment in the clinic. While some argue that continuous feedback resembles their interaction with clinic supervisors, others say that the current feedback most closely resembles real-life situations in clinics. Despite these conflicting viewpoints, and after thorough consideration, all residents concluded that continuous feedback and goal setting are essential in establishing an optimal learning environment resembling real-life practice. Residents further argued that basic professional knowledge, such as supervision by an expert, should be offered at the onset of every training session to help residents obtain a basic understanding of certain terms and procedures; 'Optimally, there could be a specialist who could help you get started. The first hour so you could get the basics. You could ask: Is this an ovary? Is this a good enough image of the uterus?' (P1.2).

When asked to reflect on how the current feedback supports residents in their learning, residents pointed to feedback as a need to resemble real-life practice by

emphasising the need for person-to-person communication; 'It is equal to the clinic if you have a supervisor who can see "well, you start with the wrong image". Then, they help to get the right image.' (P1.4). Other residents highlighted the need for clearer explanations of correct approaches to successful task completion through visual indicators; 'And then it [the simulator] says that you are 5 or 4 millimetres from where you should have measured. There, it could well have put it [the correct measurements] on the image to show the right thing.' (P1.1).

Finally, residents negatively commented on the fact that the feedback is presented at the end of the training session, arguing that smaller intervals of training through repeating tasks could result in better training outcomes. 'The shorter the intervals, the easier it is to remember and get it done. Getting it approved that it was right, and then you can move on' (P1.4). Similarly, several residents expressed an interest in returning to a specific task and redoing them to understand their mistakes. This is not possible without redoing the entire module.

> […] it could be nice to take a step back instead of having to redo it all. Taking a step back, so "okay, it was a little annoying that it [the simulator] said I had the wrong measurement. I step back and do it again, and then you could go on." (P1.3)

4.4. Expert perspectives

Given that our participants were all residents during the interviews and observations, they may not have obtained sufficient experience to fully reflect on the impact of changes to their training process. As such, we consider it critical to reflect on our findings with experienced medical experts. In this subsection, we focus on the results of our interviews with four medical experts. The expert perspectives on the insights obtained from the residents are presented below.

4.4.1. Continuous guidance during practice

The medical experts stressed that continuous feedback during training is highly essential in real-life practice, increasing the residents' confidence and ability to perform as individual clinicians. They furthermore agreed that simulation-based practice should similarly uphold the notion of continuous feedback during training. All four experts stressed that guiding residents during the procedures is the most beneficial approach to teaching; 'The most optimal is continuous feedback so that you are constantly corrected. You have some supervisor, someone who stands next to you. It provides much faster training and self-confidence for young doctors, who become independent faster.' (P2.1).

When asked whether continuous feedback resembles the feedback strategies provided in real-life training, all experts expressed the desire for continuous feedback to be an integrated part of teaching residents in real-life scenarios. However, due to the lack of resources in clinics, residents' feedback is often provided after each training session. Some stated it is far from the optimal approach in teaching and learning. Still, due to a lack of resources, there is no time to teach the residents correctly:

> Most often, due to lack of resources, it ends with continuous feedback being removed, and then you end up with feedback when you are done with it. "You did not do so well. Do it over again." You could say that it is quite far from what we know works well, continuous feedback and peer training. People are so busy that there are simply no resources to care for the young doctors even though it pays off because they become independent and skilled rapidly (P2.1).

Related to this, the majority of experts stated that continuous feedback is vital at the beginning of the resident's training as it helps build responsibility along the way – something that cannot be achieved by simply looking at an expert carrying out a given task. One of the experts describes live feedback as follows;

> This is where you go in, and tell them what to do. And then the student or the young doctor gets more and more responsibility along the way while I stand and look. Suddenly, they become independent after some time, and then they can call if trouble arises (P2.1).

The need for debriefing after each training session was stressed by some of the experts as highly critical since both supervisor and resident need to reflect upon the choices made throughout the training and discuss different implications for further improvements:

> It [continuous feedback] can not stand alone, so you also need debriefing eventually. You may need to have continuous feedback for a period of time and then be let loose and act upon your thoughts, and in debriefing afterwards, you can compare the two processes. The continuous feedback, independent period and debriefing at the end (P2.3).

Some experts further distinguished between residents' experience level and the feedback required during training sessions, highlighting that some residents need additional assistance, whereas others need less.

> It is very different in how much detail you provide every time depending on the situation and which doctor is present. Some [supervisors] also take over the training session, while others let them [residents] do it themselves and then come up with some feedback along the way (P2.2).

We found, perhaps not unexpectedly, that the differentiation in teaching and feedback provided during training sessions is dependent upon a risk assessment of patient harm, which may result in experts taking control of the procedure;

> If it is starting to get dangerous, then I'm taking over. Sometimes, people go so far into their zone that they are unable to listen. Or do not perceive that it is not fast enough because they are so concentrated. Then you have to grab the instruments and take over for a moment (P2.1).

4.4.2. Repetition of tasks

Based on the findings from our contextual inquiry, residents stressed the importance of repeating tasks. In particular, the experts highlighted the need for retrospectively selecting specific tasks for repetition instead of repeating the entire segment – an annoyance currently experienced in the simulation system. We again asked experts to reflect upon this to hear their opinions regarding repetition as an essential part of learning. Our findings showed that experts similarly see the repetition of tasks as a necessary part of learning due to enhancement of expertise:

> I completely agree with that. This is how I often feel myself. If there is something I fail at or could do better, then often I think, "What can I do." If you repeat it, then you be aware of your mistakes (P2.3).

We followed up on repetition with the experts, during which they stressed that the repeated tasks should be dynamic and change to provide different scenarios for the residents. Several experts recommend against identical repeated tasks, as it often results in memorisation instead of problem-solving, which is essential in health care. The residents need to be challenged on the topic of the exercise as opposed to challenging their ability to memorise specific procedures. These experts highlighted that although residents have the skills to memorise, they must be trained to solve problems independently. The residents need to be prepared for emergencies and how to act when they occur;

> I do not think it should be identical because then it becomes memorisation instead of problem-solving. Another task of the same character or the same type where they are challenged a bit. That would be nice because if they learn it and can do it, they will likely do it in the future. If it is just the same task, then there is a risk that it will be memorised (P2.1).

One expert suggested considering identical and non-identical tasks in repetition to see a pattern of mistakes made and then act upon these mistakes:

> It would be nice if you can do both because then you can see if there is a pattern in the mistakes they make. If they make the same type of mistake, then you know where to turn to and make that person better (P2.3).

4.4.3. AI-driven feedback based on resident experience level

The residents expressed differences in their preferences for navigating the learning material and, based on our interviews with the experts, received individually adjusted feedback during their training in accordance with their experience level and other personal characteristics. The experts expressed that residents with little to no experience can often be sensitive to receiving feedback, so the presented feedback should be diplomatic and pedagogical;

> You can say that the young doctors are more fragile, and you have to say things more diplomatically not to make them sad. There is an enormous amount of professionalism in calling yourself a doctor, so it can feel like a personal attack if someone says something about your work (P2.1).

Related to this, the experts explained that inexperienced residents receive more supervision as compared to their more experienced peers.

Our discussion on implementing AI-driven feedback based on the residents' experience level led the experts to consider the presentation of the assessment upon task completion. We learned from the residents that the simulators' current tasks assessment and feedback were too rigid and inflexible, with minor mistakes leading to failed tasks. All four experts argue that the assessment should depend on the individual resident: 'Someone who have done well throughout their education and may not have met critique,'. Our expert participants further argue that rather than solely indicating that a mistake was made, showcasing the correct or alternative procedure is valuable. One of the experts highlighted different nuances in presenting feedback as a consideration for a more pedagogical approach;

> After all, there are many ways to give feedback from such a program. There are many more nuances of feedback than just telling residents they were wrong or failed a segment. Instead, tell them that you were close to doing it right, and if it was just a millimetre away, it had been perfect (P2.1).

In addition, some experts highlighted the need for more specific and alternative assessments for completing tasks and the ability to render the procedure prior to practice to decrease mistakes and provide a more guided-practice approach to inexperienced residents.

5. Recommendations for AI-driven training feedback

Our findings provide insights into residents' expectations of optimisation within feedback in medical simulation training. Within these findings, AI support is a focal point for optimisations in future medical simulation training, given the residents' need for optimised explanations, the resemblance of real-life practice, and the correctness of in-the-moment support. Overall, the residents stressed the importance of virtual supervision to guide them through their training. Combining these insights sums up the role of AI support systems and how AI can enhance performance (Schneiderman 2022) and support users with in-the-moment support.

To provide a more concrete overview of our findings, we outline recommendations for AI feedback and explanations in the context of (clinical) simulation-based training. We summarised the experts' and medical residents' responses based on the established HCAI perspective of 'Human in the group; Computers in the loop', stating that humans are social beings, and technology should act as a collaborator to support their performance (Schneiderman 2022). Following this outlook, our findings from the exploration and validation stage were grouped into relevant themes that outline six recommendations for AI feedback and explanations in the context of simulation-based training.

Recommendation 1: Provide continuous feedback and dialogue throughout each practice session to guide and reassure residents in their training During practice, live guidance enables residents to recognise and respond to errors made as they happen. It is important to consider that continuous feedback should be non-intrusive and not 'give away' the answer. This ensures a leading role for the residents and that the AI-driven support system acts similarly to a human instructor. Current interaction with AI-driven support systems is typically turn-based, in which user input and AI recommendation follow each other in subsequent order (van Berkel et al. 2021). This paradigm of a turn-taking process lacks the opportunity for 'in-the-moment' support, as it requires the user to recognise when they need help and actively request it. However, prior work on clinical decision support tools stresses that medical practitioners are unlikely to recognise when they might need help (Yang, Steinfeld, and Zimmerman 2019). This might be due to various aspects, such as incorporated habits, wrongful guidance, or uncertainty. Therefore, the need for continuous guidance, explanations, and feedback loops is beneficial for the residents' learning. Yilmaz et al. recently stressed the relevance of continuous assessment and guidance in the operating room, pointing to the fact that most surgical skills learning happens through the continuous evaluation of resident performance by a senior instructor (Yilmaz et al. 2022).

Recommendation 2: Provide pervasive AI support at training onset, slowly subside to act as a collaborator The AI should guide the start of a training session by clearly explaining the goal and an overview of the training process. The need for the AI to be pervasive initially and then slowly subside as the residents' confidence grows aligns with the experts' perspective towards the need for residents to be independent in performing procedures. The AI feedback must not interrupt residents but instead assist them in their learning process. The need for feedback to be an integrated part of the experience in a medical workflow is discussed by Yang et al., who stress the importance of designing clinical decision-support tools to act as an integrated experience in the medical workflow (Yang, Steinfeld, and Zimmerman 2019).

Recommendation 3: Feedback must resemble the supervisor-resident interaction in real-life clinical settings Simulators are used to mimic real-world patient encounters. While residents were generally positive about the medical scenarios' realism, the feedback must resemble a virtual supervisor and support the realism of real-life clinic interaction between supervisors and residents. In clinics, however, time is paramount to treatment success, often preventing more extensive practice opportunities. Work within human-AI collaboration has showcased that users want to take the lead during interaction with AI-driven systems, emphasising that humans should be in charge of making decisions (Oh et al. 2018). However, these results do not necessarily hold within learning, given the role of the AI as a supervisor.

Recommendation 4: Feedback should be transparent to enable the residents to act upon it The degree of explainability of AI systems plays a significant role in users' trust towards these systems and their perceived suitability for collaboration (Zhang et al. 2021). We stress the importance of transparent and explainable feedback in training contexts to allow residents to understand the provided output. Prior work has highlighted that by offering explanations, AI systems can increase the user's understanding of the system (Zhang et al. 2021). Similarly, transparency and trust are vital to a successful connection between teacher and student (Mirchi et al. 2020). Therefore, we urge transparency in assessment criteria and feedback, providing the residents with actionable suggestions for improvement.

Recommendation 5: Provide retrospective access to failed tasks Repeating previously failed tasks can be beneficial in a learning environment to reduce future errors. In addition, we found that the simulator should

introduce variation in the repeated tasks rather than presenting an identical context on each trial. This ensures that residents increase their problem-solving skills instead of memorisation skills. This recommendation further aligns with prior work on spaced repetition (Pashler et al. 2007), a learning technique in which learners face tasks where they have made mistakes more frequently than tasks they completed with ease.

Recommendation 6: Align tasks and feedback to the residents' skills and experience Our insights highlight the need for tasks in simulation-based training to match residents' experience levels. Furthermore, the experts stressed the need to adjust feedback to the experience level of the residents rather than presenting identical feedback to all skill levels. Based on this, we recommend creating a dynamic task flow that intelligently aligns with the residents' skill level, incorporating the aforementioned spaced repetition approach to ensure that residents address their weaknesses.

6. Discussion

With the vast and growing global demand for reliable healthcare, adequate training of clinicians is critical. This is reinforced by the notion that diagnostic errors continue to range up to 17% in adverse events and 10% resulting in patient death (Bordini, Stephany, and Kliegman 2017). Such diagnostic errors can have disastrous consequences, leading to incorrect diagnostics and irreversible patient harm. Medical simulation-based training has been identified as a cost-efficient solution that can contribute to residents' learning by providing a structured training process. With the increasing technological advancements in healthcare, the use of AI support has been suggested for several application domains, including the incorporation of clinical decision-support tools for the implantation of artificial hearts (Cai, Reif et al. 2019; Yang, Steinfeld, and Zimmerman 2019), embedded in treatment decision-making tools (Yang, Steinfeld, and Zimmerman 2019), and as a support tool in the assessment of live video feeds (van Berkel et al. 2021).

In this work, we sought to understand through a user-centred research approach how medical residents' needs and expectations during simulation-based training can be optimised to support medical residents' learning process. Moreover, how these needs can be articulated to support the role of AI feedback in future medical simulation-based training. Our goal is to understand how the feedback presented to residents can be optimised, considering the role AI holds for future optimisation to align with the medical residents' learning process more closely. In addition to observing and interviewing medical residents, we have validated our findings with medical experts to ensure their contextual validity and alignment with the residents' learning outcomes.

6.1. AI-driven feedback in medical training

Through our contextual inquiry, we found that the feedback between residents and the simulator currently follows a turn-taking process. This often left residents guessing to identify the correct approach towards completing their training tasks, sometimes even resulting in a full stop in their training process. As summarised in Recommendation 1, we identified the need for in-the-moment support throughout the training. This observation is supported by residents' need for AI-driven feedback to closely resemble the interaction and conversation between supervisors and students in real-life practice.

The residents' desire for supervision and feedback to resemble a real-life supervisor, as highlighted in Recommendation 3, raises questions about the design of AI-driven feedback. Prior work on AI-driven feedback in learning environments stresses the importance of AI-driven systems closely resembling a human supervisor due to the increased realism and effectiveness (Mirchi et al. 2020; Salvetti et al. 2021). For AI systems to take up the role of supervisors and mimic human behaviour, the AI system needs to obtain a higher awareness of the student. This may require more accurate modelling of human behaviour (Mohan, Venkatakrishnan, and Hartzler 2020) and a transition towards more human-like collaborative approaches (Kambhampati 2019). Our results highlighted that real-life training scenarios aim to transfer skills, and in-the-moment support increases residents' confidence and abilities to perform individually.

The move from human-to-human communication to human-AI communication results in a shift in trust, as residents now need to trust the assessment of AI-driven systems instead of relying solely on the proven expertise of their senior supervisors. Especially in learning environments, transparency and trust are vital components to ensure a successful connection between teacher and student (Mirchi et al. 2020). This aligns with our findings, in which the residents expressed the need to understand the criteria for assessing a given procedure as positive or negative, as described in Recommendation 4. Research using AI to train communication skills in healthcare stresses the importance of a learner-centred approach that allows learners to receive authentic feedback in the required context

and explore different techniques in repeated encounters (Butow and Hoque 2020). Our findings from the interviewed medical experts highlight the need to provide non-identical scenarios during task repetition, as repetition of tasks can result in training residents' memorisation skills rather than the training of problem-solving skills, as pointed to in Recommendation 5. The experts highlighted that residents need expertise in handling unexpected procedures to face acute emergencies.

6.2. Designing for human-AI collaboration in learning

Our insights highlight the need for AI-driven feedback to be omnipresent at the beginning of practice, slowly diminishing and taking up the role of a collaborator as the resident's skill level increases. As summarised in Recommendation 2, the medical experts highlighted that residents acquire their independence throughout the training, with residents' responsibility towards engaging with patients growing over time as their skill level and confidence increase.

Prior research indicates that users should take the leading role in human-AI collaboration scenarios (Oh et al. 2018), whereas other research highlights conflicting needs as to whether users should be provided with detailed instructions, such as dialogue and alerts during interactions (Preece, Rogers, and Sharp 2015). From the medical simulation training context, detailed instructions and dialogue are essential to accommodate learning. The simulator might check whether a procedure is completed correctly, e.g. the probe is in the correct position. Still, it does not assess whether the resident has learned what to look for to obtain the learning standards. Several residents expressed a need for dialogue and instructions to prevent errors and incorporate bad habits. It is critical to carefully consider the timing of the feedback presented during interactions, as the feedback should not prevent the resident from figuring out the correct solutions on their own, see Recommendation 1. As confirmed by residents and experts, continuous feedback and dialogue during training are critical. Here, we argue for an AI-supported simulation system only to intervene and guide the residents after several unsuccessful attempts, as is typical in real-life clinical supervision. This prevents the residents from going down a negative feedback loop, something that we observed across the majority of residents.

AI-driven feedback during medical simulation training needs to align with the medical residents' experience levels by adjusting the assessment, scenarios, and tasks to the residents' experience levels, see Recommendation 6. Our discussion with medical experts stressed the importance of adjusting the feedback to the resident, especially regarding feedback strategy, to accommodate the residents' individual experience levels. This is because, as highlighted by the experts, novice residents are more sensitive towards their assessment, as they often equal their subject knowledge to their personality. While related research in the area of human-AI collaboration has set out to explore the spectrum of coordination, communication, and goal-sharing between humans and AI systems (McCormack et al. 2019; Oh et al. 2018; Schaekermann et al. 2020; Yang et al. 2020; Zhang et al. 2021), our work set out to understand collaboration in learning between medical residents and experts, and provided new perspectives in the area for AI-driven training feedback in health care.

6.3. Explainability in a learning context

With the rising demand for AI support in the high-stake healthcare sector, extensive work has focused on making AI systems more interpretive, transparent, and explainable. This is critical for the adoption of AI-based technology, as previous studies have shown that individuals are more reluctant to adopt technologies that are not interpretive and trustworthy (Arrieta et al. 2020). Despite the widely shared notion that explainability is essential, consensus on 'what makes AI explainable?' is lacking. The meaning and value of explainability and related concepts such as transparency and trustworthiness differ between individuals. Arguably, the key factor in implementing AI in health care is to make the output of intelligent systems 'understandable to the knowledge of all kinds of users' (Adadi and Berrada 2020), including e.g. nurses, doctors, and administration. Furthermore, we stress that different contexts lead to different requirements and expectations for explainable AI.

The learning context demands a different interpretation and implementation of explainability compared to other contexts since explanations can adversely affect the learning process. For example, instantly revealing correct answers without accommodating the learner's opportunity to reflect or diminishing a learner's motivation by continuously pointing to errors are examples of how that could manifest. With this, it is important to consider the interplay between learners and AI and the role AI holds in educational purposes in providing augmentation that emphasises humans and AI as equal team members (Molenaar 2022). The level of automation of AI in educational environments is tied to the relations of control between the learner and the AI and the challenge in AI systems to intelligently adapt their presentation to the task context of the user (van Berkel et al. 2023).

Within a training context, the end-users may not necessarily represent different roles – but they are likely to represent different skill levels and educational progress. Therefore, the automation of AI in this context will need to adapt the frequency, level of detail, and support provided in the explanations to the user's skill level. Within this interplay, providing a two-directional interaction between AI and learners can contribute to equal collaborations and minimise the gap between full AI automation and complete human control (Molenaar 2022). Considering the levels of AI automation, our insights position automation between high and partial control. Given that high AI, automation is required at the beginning of the interaction to guide the learner and then slowly subsides to act as a collaborator, see recommendation-2, providing more partial automation through turn-taking processes (Molenaar 2022).

Further, our results show that residents doubted whether AI-based feedback could closely resemble real-life training with a supervisor, as communication is a central aspect of the training. For example, some residents explicitly highlighted the need for human-to-human communication to accommodate their learning process. The ability of an AI-based feedback system to resemble a human supervisor is partly determined by its degree of transparency, as the lack thereof can result in negative emotions from students using the technology (Mirchi et al. 2020). This aligns with our findings that residents highlighted the need for more in-depth explanations throughout the training (e.g. continuous assessments and decreasing the experience of doubt during training).

6.4. Limitations & future work

We recognise a number of limitations in our work that ought to be considered when interpreting our results. First, despite our extensive efforts in recruiting, assisted by our medical partners, our study is limited in its sample size (nine residents and four medical experts). We acknowledge the challenges of data saturation and the reporting of such since saturation is not known until it is reached (Caine 2016). Given the difficulty of reporting sufficient evidence of data saturation, we confidently base our sample size's validity on the clear identification of patterns in our data and the subsequent validation and assessment of residents' data with highly specialised medical experts.

Moreover, the sample size in our study is grounded in pragmatic considerations regarding the availability of the participants (Vasileiou et al. 2018). Challenges in recruiting medical experts for study participation are well-known due to required specialist knowledge in combination with tight and unpredictable schedules (van Berkel et al. 2021; VanGeest, Johnson, and Welch 2007).

Second, while the residents we observed had different specialisation backgrounds (e.g. general practice, gynaecology), they all used the transvaginal ultrasound simulator. As a result, our observations are limited to the tasks encountered within gynaecology and ultrasound. Therefore, this observation study did not consider other procedures, e.g. surgery-related tasks. To broaden the scope of our insights, we aimed to reflect on simulation-based training more broadly in our interviews with the residents and medical experts. Despite these efforts, the presented results should be extended through studies focusing on other medical disciplines.

As a part of our data collection, we relied specifically on our semi-structured interviews and observations with the residents to understand the limitations of simulation-based training. To broaden the scope of our data collection for future considerations, the reports generated by the simulation would be beneficial to include to provide more detailed data on challenges arising in simulation-based training.

Third, the simulator used in this study provides a limited amount of feedback and interactivity. As such, the simulators act as a highly rigid practice tool for predefined tasks, not necessarily providing the flexibility required to increase residents' learning. We see an opportunity for future work to investigate the effect of continuous feedback as used in more sophisticated AI-backed systems. Finally, our recruitment location is limited to two hospitals in Denmark. While the curriculum on obstetrics and gynaecology is standardised mainly across Europe (van der Aa et al. 2016), significant differences in training needs, expectations, and medical culture may exist between countries and hospital units.

7. Conclusion

In this work, we set out to understand the needs of clinical residents in their use of medical training simulators and their expectations and wishes for AI feedback and explanations in the context of clinical learning. Our interviews and contextual inquiries, in which we involve nine medical residents and four medical experts, highlight that while residents had multiple and different motivations for using the simulator, they unanimously agreed that the feedback currently provided by the simulator is too rigid and is not presented when it is most beneficial. Rather than waiting for a training assessment to finish, residents highlight the need for more continuous feedback throughout the training procedure, and this aligns with the medical experts' perspective on feedback during medical training. Our

results show that residents experience frustrations in using simulation-based training due to the contrast between feedback provided by the simulator and the rich explanations and live feedback as experienced during real-life medical practice under the supervision of a human expert. To support the extensive opportunities medical simulators provide to train residents efficiently and without inducing patient harm, we present six recommendations for developing AI-driven training feedback in the context of medical simulators. We discuss the role of feedback in (medical) training scenarios as an under-explored but highly relevant area of explainable AI. Explanations are at the core of teaching and learning and, therefore, provide a golden opportunity for the HCI and AI research communities to develop engaging and realistic training resources that align with the wishes and needs of learners.

Note

1. https://www.intelligentultrasound.com/.

Acknowledgments

The authors would like to thank all the residents and clinicians who took part in the studies for their valuable contributions.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work is supported by Digital Research Centre Denmark (DIREC) project 'Explain-Me' under Innovation Fund Denmark.

ORCID

Niels van Berkel http://orcid.org/0000-0001-5106-7692

References

Adadi, Amina, and Mohammed Berrada. 2020. "Explainable AI for Healthcare: From Black Box to Interpretable Models." In *Embedded Systems and Artificial Intelligence*, edited by Vikrant Bhateja, Suresh Chandra Satapathy, and Hassan Satori, 327–337. Singapore: Springer Singapore. https://doi.org/10.1007/978-981-15-0947-6_31. ISBN: 978-981-15-0947-6.

Amir, Ofra, and Ya'akov (Kobi) Gal. 2013. "Plan Recognition and Visualization in Exploratory Learning Environments." *ACM Transactions on Interactive Intelligent Systems* 3 (3): Article 16, 23 pages. https://doi.org/10.1145/2533670.2533674.

Arrieta, Alejandro Barredo, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, et al. 2020. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI." *Information Fusion*58:82–115. https://doi.org/10.1016/j.inffus.2019.12.012.

Bellio, Maura, Dominic Furniss, Neil P. Oxtoby, Sara Garbarino, Nicholas C. Firth, Annemie Ribbens, Daniel C. Alexander, and Ann Blandford. 2021. "Opportunities and Barriers for Adoption of a Decision-Support Tool for Alzheimer's Disease." *ACM Transactions on Computing for Healthcare* 2 (4): Article 32, 19 pages. https://doi.org/10.1145/3462764.

Bjork, Robert. 1994. *Memory and Meta-Memory Considerations in the Training of Human Beings*, 185–205. 255 Main Street, 9th Floor Cambridge, MIT Press. https://doi.org/10.7551/mitpress/4561.003.0011.

Bjork, Elizabeth Ligon, and Robert A. Bjork. 2011. *Making Things Hard on Yourself, But in a Good Way: Creating Desirable Difficulties to Enhance Learning*, 56–64. New York, NY, US, Worth Publishers. https://api.semanticscholar.org/CorpusID:894160.

Bordini, Brett J., Alyssa Stephany, and Robert Kliegman. 2017. "Overcoming Diagnostic Errors in Medical Practice." *The Journal of Pediatrics* 185:19–25.e1. https://doi.org/10.1016/j.jpeds.2017.02.065.

Braun, Virginia, and Victoria Clarke. 2022. *Thematic Analysis*. Sage, London.

Butow, Phyllis, and Ehsan Hoque. 2020. "Using Artificial Intelligence to Analyse and Teach Communication in Healthcare." *The Breast* 50:49–55. https://doi.org/10.1016/j.breast.2020.01.008.

Cai, Carrie J., Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, et al. 2019. "Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making." In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*, 1–14. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3290605.3300234. ISBN: 9781450359702.

Cai, Carrie J., Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. ""Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making." *Proceedings of the ACM on Human-Computer Interaction* 3 (CSCW): Article 104, 24 pages. https://doi.org/10.1145/3359206.

Caine, Kelly. 2016. "Local Standards for Sample Size at CHI." In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*, 981–992. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2858036.2858498. ISBN: 9781450333627.

Croskerry, P., and G. R. Nimmo. 2011. "Better Clinical Decision Making and Reducing Diagnostic Error." *The Journal of the Royal College of Physicians of Edinburgh* 41 (2): 155–162. https://doi.org/10.4997/JRCPE.2011.208.

Došilović, Filip Karlo, Mario Brčić, and Nikica Hlupić. 2018. "Explainable Artificial Intelligence: A Survey." In *2018 41st International Convention on Information and*

Communication Technology, Electronics and Microelectronics (MIPRO), 0210–0215. https://doi.org/10.23919/MIPRO.2018.8400040.

Dyre, Liv, Ann Tabor, Charlotte Ringsted, and Martin G. Tolsgaard. 2017. "Imperfect Practice Makes Perfect: Error Management Training Improves Transfer of Learning." *Medical Education* 51 (2): 196–206. https://doi.org/10.1111/medu.13208.

Fitzpatrick, Geraldine, and Gunnar Ellingsen. 2013. "A Review of 25 Years of CSCW Research in Healthcare: Contributions, Challenges and Future Agendas." *Computer Supported Cooperative Work (CSCW)* 22 (4–6): 609–665. https://doi.org/10.1007/s10606-012-9168-0.

Galvin, Shelley L., and Elizabeth Buys. 2012. "Resident Perceptions of Service Versus Clinical Education." *Journal of Graduate Medical Education* 4 (4): 472–478. https://doi.org/10.4300/JGME-D-11-00170.1.

Gordon, J. A., W. M. Wilkerson, D. W. Shaffer, and E. G. Armstrong. 2001. "'Practicing' Medicine Without Risk: Students' and Educators' Responses to High-fidelity Patient Simulation." *Academic Medicine* 76 (5): 469–472. https://doi.org/10.1097/00001888-200105000-00019.

Grudin, Jonathan. 1988. "Why CSCW Applications Fail: Problems in the Design and Evaluationof Organizational Interfaces." In *Proceedings of the 1988 ACM Conference on Computer-Supported Cooperative Work (CSCW '88)*, 85–93. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/62266.62273. ISBN: 0897912829.

Juanes-Mendez, Juan A., Amaia Yurrebaso Macho, Raquel Guzmán-Ordaz, Eva Picado-Valverde, and Alexander L. Ward Mayens. 2021. "Methodology for Learning and Acquiring Clinical Skills Through Simulation with Artificial Human Models." In *Ninth International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM'21) (TEEM'21)*, 274–278. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3486011.3486461. ISBN: 9781450390668.

Kambhampati, Subbarao. 2019. "Challenges of Human-Aware AI Systems." *AI Magazine*, 41(3). 3–17. https://doi.org/10.1609/aimag.v41i3.5257.

Kneebone, R. L., W. Scott, A. Darzi, and M. Horrocks. 2004. "Simulation and Clinical Practice: Strengthening the Relationship." *Medical Education* 38 (10): 1095–1102. https://doi.org/10.1111/j.1365-2929.2004.01959.x.

Ma, Shuai, Ying Lei, Xinru Wang, Chengbo Zheng, Chuhan Shi, Ming Yin, and Xiaojuan Ma. 2023. "Who Should I Trust: AI or Myself? Leveraging Human and AI Correctness Likelihood to Promote Appropriate Trust in AI-Assisted Decision-Making." In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*, Article 759, 19 pages. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3544548.3581058. ISBN: 9781450394215.

McCormack, Jon, Toby Gifford, Patrick Hutchings, Maria Teresa Llano Rodriguez, Matthew Yee-King, and Mark d'Inverno. 2019. "In a Silent Way: Communication Between AI and Improvising Musicians Beyond Sound." In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*, 1–11. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3290605.3300268. ISBN: 9781450359702.

Mirchi, Nykan, Vincent Bissonnette, Recai Yilmaz, Nicole Ledwos, Alexander Winkler-Schwartz, and Rolando F. Del Maestro. 2020. "The Virtual Operative Assistant: An Explainable Artificial Intelligence Tool for Simulation-based Training in Surgery and Medicine." *PLoS ONE* 15 (2): 1–15. https://doi.org/10.1371/journal.pone.0229596.

Mohan, Shiwali, Anusha Venkatakrishnan, and Andrea L. Hartzler. 2020. "Designing An AI Health Coach and Studying Its Utility in Promoting Regular Aerobic Exercise." *ACM Transactions on Interactive Intelligent Systems* 10 (2): Article 14, 30 pages. https://doi.org/10.1145/3366501.

Mohseni, Sina, Niloofar Zarei, and Eric D. Ragan. 2021. "A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems." *ACM Transactions on Interactive Intelligent Systems* 11 (3–4): Article 24, 45 pages. https://doi.org/10.1145/3387166.

Molenaar, Inge. 2022. "Towards Hybrid Human-AI Learning Technologies." *European Journal of Education* 57 (4): 632–645. https://doi.org/10.1111/ejed.12527.

Musen, Mark A., Blackford Middleton, and Robert A. Greenes. 2014. *Clinical Decision-Support Systems*, 643–674. London: Springer London. https://doi.org/10.1007/978-1-4471-4474-8_22.ISBN: 978-1-4471-4474-8.

Oh, Changhoon, Jungwoo Song, Jinhan Choi, Seonghyeon Kim, Sungwoo Lee, and Bongwon Suh. 2018. "I Lead, You Help but Only with Enough Details: Understanding User Experience of Co-Creation with Artificial Intelligence." In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*, 1–13. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3173574.3174223. ISBN: 9781450356206.

Okamura, Kazuo, and Seiji Yamada. 2020. "Adaptive Trust Calibration for Human-AI Collaboration." *PLoS ONE* 15: Article e0229132. https://doi.org/10.1371/journal.pone.0229132.

Panigutti, Cecilia, Andrea Beretta, Fosca Giannotti, and Dino Pedreschi. 2022. "Understanding the Impact of Explanations on Advice-Taking: A User Study for AI-Based Clinical Decision Support Systems." In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*, Article 568, 9 pages. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3491102.3502104. ISBN: 9781450391573.

Pashler, Harold, Doug Rohrer, Nicholas J. Cepeda, and Shana K. Carpenter. 2007. "Enhancing Learning and Retarding Forgetting: Choices and Consequences." *Psychonomic Bulletin & Review* 14 (2): 187–193. https://doi.org/10.3758/BF03194050.

Preece, Jennifer, Yvonne Rogers, and Helen Sharp. 2015. *Interaction Design: Beyond Human–Computer Interaction*. Hoboken, NJ: Wiley. ISBN: 978-1-119-02075-2.

Raven, Mary Elizabeth, and Alicia Flanders. 1996. "Using Contextual Inquiry to Learn About Your Audiences." *ACM SIGDOC Asterisk Journal of Computer Documentation* 20 (1): 1–13. https://doi.org/10.1145/227614.227615.

Salas, Eduardo, Katherine A. Wilson, C. Shawn Burke, and Heather A. Priest. 2005. "Using Simulation-Based Training to Improve Patient Safety: What Does It Take?." *The Joint Commission Journal on Quality and Patient Safety* 31 (7): 363–371. https://doi.org/10.1016/S1553-7250(05)31049-X.

Salvetti, Fernando, Roxane Gardner, Rebecca D. Minehart, and Barbara Bertagni. 2021. "Enhanced Reality for Healthcare Simulation." In *Recent Advances in Technologies for Inclusive Well-Being*, 103–140. Switzerland AG: Springer International Publishing,: https://doi.org/10.1007/978-3-030-59608-8_7.

Schaekermann, Mike, Graeme Beaton, Elaheh Sanoubari, Andrew Lim, Kate Larson, and Edith Law. 2020. "Ambiguity-aware AI Assistants for Medical Data Analysis." In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*, 1–14. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3313831.3376506. ISBN: 9781450367080.

Schneiderman, Ben. 2022. *Human-Centered AI*. Great Clarendon Street, Oxford, OX2, 6DP, United Kingdom: Oxford University Press. ISBN: 9780192845290.

Springer, Aaron, and Steve Whittaker. 2020. "Progressive Disclosure: When, Why, and How Do Users Want Algorithmic Transparency Information?." *ACM Transactions on Interactive Intelligent Systems*10 (4): Article 29, 32 pages. https://doi.org/10.1145/3374218.

van Berkel, Niels, Omer F. Ahmad, Danail Stoyanov, Laurence Lovat, and Ann Blandford. 2021. "Designing Visual Markers for Continuous Artificial Intelligence Support: A Colonoscopy Case Study." *ACM Transactions on Computing for Healthcare* 2 (1): Article 7, 24 pages. https://doi.org/10.1145/3422156.

van Berkel, Niels, Maura Bellio, Mikael B. Skov, and Ann Blandford. 2023. "Measurements, Algorithms, and Presentations of Reality: Framing Interactions with AI-Enabled Decision Support." *ACM Transactions on Computer-Human Interaction* 30 (2): Article 32, 33 pages. https://doi.org/10.1145/3571815.

van Berkel, Niels, Matthew J. Clarkson, Guofang Xiao, Eren Dursun, Moustafa Allam, Brian R. Davidson, and Ann Blandford. 2020. "Dimensions of Ecological Validity for Usability Evaluations in Clinical Settings." *Journal of Biomedical Informatics* 110:Article 103553. https://doi.org/10.1016/j.jbi.2020.103553.

van der Aa, Jessica E., Angelique J. Goverde, Pim W. Teunissen, and Fedde Scheele. 2016. "Paving the Road for a European Postgraduate Training Curriculum." *European Journal of Obstetrics & Gynecology and Reproductive Biology* 203:229–231. https://doi.org/10.1016/j.ejogrb.2016.05.020.

VanGeest, Jonathan B., Timothy P. Johnson, and Verna L. Welch. 2007. "Methodologies for Improving Response Rates in Surveys of Physicians: A Systematic Review." *Evaluation & the Health Professions*30 (4): 303–321. https://doi.org/10.1177/0163278707307899.

Vasileiou, Konstantina, Julie Barnett, Susan Thorpe, and Terry Young. 2018. "Characterising and Justifying Sample Size Sufficiency in Interview-based Studies: Systematic Analysis of Qualitative Health Research Over a 15-Year Period." *BMC Medical Research Methodology* 18 (1): 148. https://doi.org/10.1186/s12874-018-0594-7.

Wang, Dakuo, Elizabeth Churchill, Pattie Maes, Xiangmin Fan, Ben Shneiderman, Yuanchun Shi, and Qianying Wang. 2020. "From Human–Human Collaboration to Human–AI Collaboration: Designing AI Systems That Can Work Together with People." In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (CHI EA '20)*, 1–6. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3334480.3381069. ISBN: 9781450368193.

Wang, Dakuo, Liuping Wang, Zhan Zhang, Ding Wang, Haiyi Zhu, Yvonne Gao, Xiangmin Fan, and Feng Tian. 2021. "'Brilliant AI Doctor' in Rural Clinics: Challenges in AI-Powered Clinical Decision Support System Deployment." In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*, Article 697, 18 pages. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3411764.3445432. ISBN: 9781450380966.

Wolff, Margaret, Mary Jo Wagner, Stacey Poznanski, Jocelyn Schiller, and Sally Santen. 2015. "Not Another Boring Lecture: Engaging Learners with Active Learning Techniques." *The Journal of Emergency Medicine* 48 (1): 85–93. https://doi.org/10.1016/j.jemermed.2014.09.010.

Yang, Qian, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. "Re-Examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design." In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*, 1–13. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3313831.3376301. ISBN: 9781450367080.

Yang, Qian, Aaron Steinfeld, and John Zimmerman. 2019. "Unremarkable AI: Fitting Intelligent Decision Support into Critical, Clinical Decision-Making Processes." In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*, 1–11. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3290605.3300468. ISBN: 9781450359702.

Yilmaz, Recai, Alexander Winkler-Schwartz, Nykan Mirchi, Aiden Reich, Sommer Christie, Dan Huy Tran, Nicole Ledwos, et al. 2022. "Continuous Monitoring of Surgical Bimanual Expertise Using Deep Neural Networks in Virtual Reality Simulation." *NPJ Digital Medicine* 5 (1): 54. https://doi.org/10.1038/s41746-022-00596-8.

Zhang, Rui, Nathan J. McNeese, Guo Freeman, and Geoff Musick. 2021. "'An Ideal Human': Expectations of AI Teammates in Human-AI Teaming." *Proceedings of the ACM on Human-Computer Interaction*4 (CSCW3): Article 246, 25 pages. https://doi.org/10.1145/3432945.

Appendices

Appendix 1. Interview guide – residents

A.1. Understand the environment

For this part of the interview, we will ask some questions regarding your profession, the lab, and the reason why you are using this lab.

- What is your name?
- What is your profession (student, doctor, professor?)
- Are you working towards a specialisation / what is your specialisation? 3. Can you tell us the reason why you are using this lab?
- What is the purpose of using this environment?
- Do you have specific learning goals in mind before entering this lab?

- You have now practised from 08:00-15:00. Can you tell us about the different tasks you set out to do and whether you were successful in achieving them?
- Why/Why not?

A.2. Training tools in action (trainee)

Within this section of the interview, we will ask you questions regarding feedback. If the simulators provided you with any feedback, how you overcame obstacles you might have encountered, and the lack of feedback provided by the simulators.

A.2.1. Experiences and feedback with the simulator.

- Can you tell us about your experience with the simulator? What went well for you? What did not go so well?
- Can you tell us about any complications during your interaction with the simulator (if any?) Were you given any feedback when you ran into these obstacles?
- What were your expectations for this feedback? Did it live up to your expectations?
- Not any feedback? What feedback would be beneficial for you to get passed these obstacles?
- Can the feedback provided be improved in any way?

A.2.2. Timestamps of specific problems. (Show different timestamps of when the trainee asked for help or indicated that he/she had trouble with the task)

- What did you experience here? (Repeat questions for every timestamp)
- How did you overcome this challenge? (What helped you overcome it?)
- Are you aware of the problem/mistakes you made, and how not to make the same mistake again?

A.3. Optimisation in learning

We have now talked about some challenges/difficulties, you have encountered during your practice with the simulator.

- Has the simulator helped you become more aware of your challenges (the ones you must work on), and your successes?
- How does this tool contribute to your learning, and how can it be optimised in relation to providing you with feedback during your practice with the simulator (including making mistakes? What went well)?
- What kind of feedback would you expect from a learning tool?
- Does the simulator you used provide this type of feedback?
- How can it be optimised so it can support medical students in practice and contribute to their learning?

A.4. Interaction and feedback

(Looking at the interaction between the student-worker and the trainee to understand how and if the feedback contributed to their learning)

- Can you talk about your experience with the student-worker in the room? a. How did the student worker help you in your training?
- How often did you get help from the student worker? a. What type of help?
- How do you learn best? Is it through continuous dialogue, debriefing etc.?

Appendix 2. Interview guide – experts

A.5. Initial understanding

- What is your full name?
- What position are you in?
- How long have you been in this position?
- What are the gaps in simulation-based training compared to training in the clinic?
- What are the primary challenges with simulation-based training compared to clinic training?
- What are the similarities?

A.6. Validation of insights

The first insight we want to highlight is 'The feedback in the ultrasound simulation'

A.6.1. Explanation of the insights regarding feedback.

The residents expressed the need for continuous feedback throughout their training. Often, when in doubt, residents guessed without knowing whether the completion was right or wrong. The lack of knowledge often led to a complete stop in training. There was a clear indication of the need for support during the training to help the residents in situations of doubt. Some residents mentioned that ongoing dialogue and feedback resembled the clinic, while others mentioned that the final feedback/debriefing was more like the clinic.

Questions for validating the insight presented above

- How do you see the aspect of ongoing feedback/dialogue as an essential element in teaching the residents?
- Is it beneficial to use this approach or do we need end feedback/debriefing to accommodate their learning?
- How does it work in the clinic? Do you intervene when the residents make a mistake or correct them before going into some challenging areas? Is it continuous or only when they ask for help?

  Our insights showed that the residents needed more precise feedback regarding the assessment for the right approach to solving tasks. Such as: 'You should have measured here instead.' 'This is not the right way to handle the probe and do this instead.'
- How do you present your feedback? Do you specify the general error or present them with alternatives?
- Would it be beneficial to incorporate this into simulation-based training? Why/why not?
- How can this feedback be designed and incorporated into the simulator? The

A.6.2. Explanation of insights regarding repetition of tasks.

Some of the residents repeated failed tasks by redoing the entire task segment. Some differentiated difficulty levels of incorrectly performed tasks by determining which tasks were important enough to redo and which were not so important. Residents emphasised the importance of retrospectively selecting tasks rather than redoing entire segments as currently found in the simulator

Questions for validating the insight presented above

- How do you see repetition of tasks as an essential element in learning? In the clinic and simulations-based environments.
- Would it be more beneficial to provide a slightly different task/context when a task is repeated when done incorrectly, or is it better to keep the task identical?
- If we look at your perspective in relation to teaching. What are your views on repetition regarding teaching? Is it a necessary feature to include in doctors' training?

A.6.3. Explanation of differences in experience levels. Our insights indicate that the residents distinguished between different ways of learning, such as see one-do one-learn one and self-directed learning. Furthermore, the residents stated the difference between being new to gynecology and having little experience to indicate how to learn most optimally. A resident mentioned that in this field, where the focus is primarily on hand maneuvers, it is essential to have continuous feedback, so that you do not strive for some bad habits.

Questions for validating the insight presented above

- What is your view on these ways of learning and how does the focus on giving feedback during practice affect teaching and learning?
- Are there different needs for beginners and more experienced doctors?
- Are different learning styles something to consider for incorporation into simulation-based training?
- How is the communication between you and the residents during training? Can you take us through the process?
- Some residents mentioned that teacher/student contact during training cannot be replaced by a system. What is your opinion towards this?

A.6.4. Explanation of 'rigid assessment of task completion'. Our insight shows that the simulation has a strict assessment of task completion. Several of the residents emphasised that the rigorous assessment was based on millimetre measurement errors, which led to errors in the given tasks.

Questions for validating the insight presented above

- What are the prerequisites for assessing a task completion as unsuccessful? Is it down to measurements, wrong image, or the scan?
- How do you assess a task as failed in practice?
- Would it be beneficial to incorporate a different assessment process if we are considering the simulation-based training?
- What should the assessment process be? Like the clinic?

A.6.5. General questions. We know that lack of time is critical in clinics. If we take time out of the equation, what would your approach to teaching residents in practice be?

- Does it correlate with what we have discussed today?
- Can you describe what the perfect simulation-based training would look like? How does it differ from what exists today?