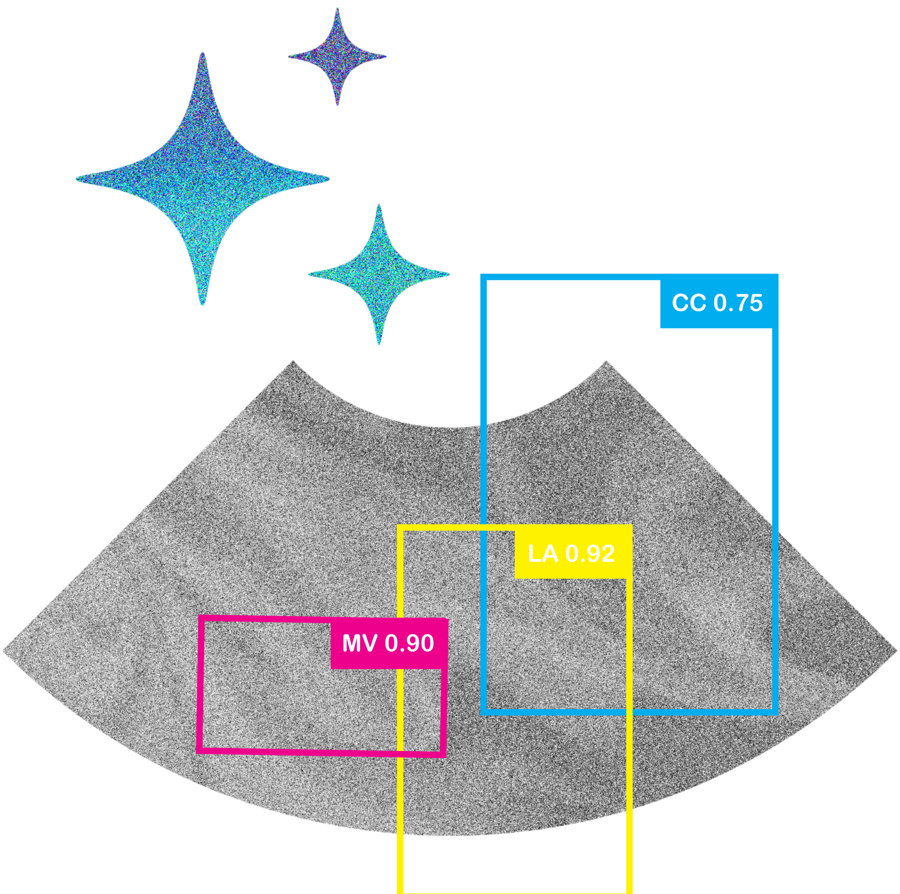
June 7th 2024

Designing Responsible AI Interfaces in Obstetric Ultrasound

Andreas Biering von Essen
&
Andreas Scott Magle Pedersen

Master's Thesis



Supervisor: Anders Hansen Henten

Acknowledgements

This master's project was made in collaboration with the SONAI-group at the Danish Pioneer Center for Artificial Intelligence in Copenhagen - P1.
We owe our deepest thanks to Morten Svendsen of CAMES (the Copenhagen Academy for Medical Simulation) and the SONAI-group for seeing the potential of the project and helping set up the first initial meetings. This gratitude extends to Martin Tolsgaard who welcomed us as affiliated master's students with the SONAI-group.
Thank you to Mary Ngo and Chun Kit Wong for facilitating many of our user-testing sessions and for countless hours of sparring and answering our questions about medical and technical details
This study would not be possible without our participants who took time out of their calendars to let their voices be heard in this project.
Lastly, we would like to thank our supervisor Anders Hansen Henten for all the useful feedback and pointers throughout the iterations of the Synopsis and the Article.

Designing Responsible AI Interfaces in Obstetric Ultrasound

Part I: Synopsis

A Techno-Anthropological Approach to Value Sensitive Design

Andreas Biering von Essen

&

Andreas Scott Magle Pedersen

AALBORG UNIVERSITY

STUDENT REPORT

Supervisor

Anders Hansen Henten

Abstract

This Master's thesis investigates the development of an Artificial Intelligence (AI) solution for the obstetric ultrasound procedure of fetal weight estimation (EFW) by a Danish research group at the Pioneer Center for AI in Copenhagen. We employ social scientific methods and newer branches of methodologies from Value Sensitive Design (VSD) to develop design requirements for a responsible User Interface (UI), with a focus on enhancing clinical workflow and ensuring compatibility with (European) societal ethics. This study finds that there are issues with the current prototype's mediation of values relating to clinical certainty and moral accountability, and proposes a shift of focus among the design team from minimizing UI elements to ensuring a robust and explicable UI that strengthens the clinician's decision-making. We present mock-up's of the UI incorporating our suggestions. By aligning specific sections of AI-development such as UI-design with societal values and stakeholder needs, researchers and designers can contribute in fostering responsible innovation and promote the effective utilization of AI in clinical settings.

Contents

I Description of submission format

This master's thesis submission consists of two parts:

1) A synopsis (127.544 characters with spaces = 53.14 pages)

2) A paper (52.912 characters with spaces = 22.04 pages)

The total submission consists of 75.28 pages. The format has been approved by the Study Board of Sustainable Design and Techno-Anthropology at the Department of Planning at Aalborg University Copenhagen. Please note that some text passages will be present in both the article and the synopsis.

It was a clear term set by the agreement with our collaborative partner, the SONAI-group at the Pioneer Center for Artificial Intelligence, that our investigation would lead to a publication. In exchange, we were granted access to facilities and areas of empirical interest, as well as countless hours of sparring with the associated researchers.

The article is ready for submission at the Journal of AI and Society following all the formal requirements for an article submission. This journal was chosen early in the process based on an extensive review of papers akin to the one we envisioned. Although a few journals scored higher on our relevancy classification, we also had to factor in our position as master's students, and AI & Society has a channel for student submissions. We have not submitted the article as per the date of thesis submission (June 7th, 2024), as the deadline with our collaborative partner lies in July, and we wanted the possibility to incorporate feedback from our master's thesis oral exam into the paper.

II About AI and Society

Our aimed-for journal of publication is AI and Society, which is an international journal established in 1987 and publishes a broad range of scholarly articles, reviews, position papers and debates along with other publications. The scope of the journal is interdisciplinarity and welcomes contributions from a wide range of fields ranging from humanities and social sciences to information technologies that centers around AI. The theme of the journal is often within societal issues including publications with a particular emphasis on cultural, social, cognitive, ethical or philosophical implications. Technological innovation holds great potential for societies, but also poses potential existential risks. One of the goals of the journal is to promote the understanding of the potential transformative impacts and critical consequences of pervasive technologies for societies. The journal is rooted in the human-centered tradition of science and technology (Submission Guidelines, AI and Society 2024).

III State of the Art

The demand for healthcare services is increasing as well as technological advancement within the sector. According to Bohr and Mermarzdeh (2020) there is already a great interest in exploring the possibilities of the applications of AI in the healthcare system that is poised to provide substantial improvement in everything from diagnostics to treatment. This includes AI that can help with the acquisition and interpretation of image-based examinations such as ultrasound, where AI has the potential to enhance the detection rates of pregnancies at risk as shown by Tolsgaard et al. (2020) and Topol (2019). One of the main talking points in the medical literature on AI for medical devices is making AI that can support clinical decision making without attempting to replace the clinicians (Bohr & Memarzadeh, 2020).
Fischer et al. (2023) conducted a qualitative study examining obstetrical clinicians' perceptions of AI as a diagnostic tool for ultrasound. The study concludes that AI solutions should enhance clinicians' decision-making capabilities rather than solely focus on clinical outcomes. Building a general professional trust in AI decisions is crucial and can be achieved through scientific validation (RCTs), professional endorsements, and ensuring explainability. Similarly, Reverberi et al. (2022) observed that foreshadowed issues such as over-reliance, under-reliance, or opaque AI judgments did not become reality amongst their participants. Clinicians developed accurate mental models of the AI's error margins by interpreting AI output cues, much like they assess a colleague's confidence. Meske & Bunde (2023) and Lieberman (2009) explores user interfaces (UI) in AI and how it can connect the user with the inner workings of the AI with actionable information. According to Corrales Compagnucci et al. (2020) effective UI design can increase comprehension for the users and help to address ethical considerations promoting responsible human-AI-collaboration. In the later years, there have been calls for investigations into the operational environment between the users and their use of the AI with a focus on acceptability, ethics and societal values (e.g. Dahlin 2021).
Collingridge (1980) has described the dilemma of control in technological developments. It can be difficult to predict and control the trajectory of a technology's effect on society in the early design stages, and when a technology is in society, it is too late to change it. Kudina & Verbeek (2019) propose an approach to deal-

ing with the Collingridge dilemma using a blend of social science and technological philosophical concepts from post-phenomenology (e.g. Ihde 1990). A well known method to incorporate values of the users into technological innovations is that of Value Sensitive Design (VSD) (e.g. Friedman et al. 2003; van de Poel & Royakkers, 2011). With the user driven methodology of VSD it is possible to identify values that are at stake in the development of technology and to materialize them into a design. Verbeek adapted VSD to properly address the Collingridge dilemma with the Values that Matter-approach (Smits et al., 2019, 2022) and Umbrello & van de Poel (2021) bring VSD into AI developments with their AI-specific approach into the field of responsible design. The aim of this master's project is to build upon these pillars of responding to calls for social scientific projects of AI-UI-design using the newest envisioned approaches of VSD.

1 Introduction

Ever since the first fire was lit, technology has changed and shaped the way we humans are in the world. One of the main results of technology from agriculture and toolmaking to industrialization is the freeing up of hands to do other work. The development of computers and information technology in the mid-20th century is no exception and automated complex calculations and data processing tasks, freeing human minds from repetitive and error-prone work. In the late 20th and early 21st centuries, the rise of the internet and digital communication tools enabled instantaneous information sharing and remote work, further reducing the need for physical presence and labor in many tasks. Now with the advent of advanced machine learning algorithms, we are at the cusp of another major shift. These algorithms, commonly referred to as artificial intelligence (AI), can not only automate a wide range of cognitive tasks that were previously thought to be the domain of humans, but also discover deep connections between data and phenomena incomprehensible to the human mind.

This master's project examines an AI model in development that is capable of assessing the quality of ultrasound images and can provide decision support for clinicians estimating fetal weight (EFW). The support from AI can limit disparities and enhance the quality of obstetric care by addressing diagnostic uncertainty and educating clinicians performing obstetric ultrasound procedures. As with all computer systems, the intuitiveness of the user interface (UI) sets the bar for the prerequisite understanding the user needs of the underlying technical workings of the system. The problem with AI however is not even the engineers of the algorithm can decipher exactly what turns the input into an output. The UI of an AI system thus has to make sense of the black boxed inner workings of the algorithms, translating complex processes into accessible and actionable insights for the end users.

1.1 Research Question

Artificial Intelligence is entering the clinic. This demands insights into user-technology dynamics at play and societal implications of AI-feedback. We had the opportunity to look into a constellation where AI is present in a clinical setting and to examine the interplay between the clinician and the AI. To ensure a sustainable AI-system that aligns with the needs of the clinic and societal values, the point of reference will center around those who have to use technology.

We are collaborating with the SONAI-group at the Pioneer Center for AI in Copenhagen, and are delivering specific design recommendations for their UI system for their AI to be used for ultrasound estimations of fetal weight. Our research question thus reflects that this thesis generates a specific output for re-design of that specific UI system.

How can the needs of healthcare practitioners engaged in fetal weight estimation using ultrasound be translated into design requirements for the development of user interfaces for AI-driven obstetric ultrasound by employing principles of Value Sensitive Design?

2 Background

The purpose of this chapter is to define all the concepts, institutions, and otherwise prerequisite knowledge for this thesis. First our working definitions and understanding of artificial intelligence (AI) will be presented at a level appropriate for the scope of this project. This will be followed by an overview of the medical practice in which the AI will be implemented. Lastly the different relevant actors and their relationships will be presented.

2.1 Artificial Intelligence

Traditionally in both the scientific and science-fiction literature, the term "Artificial Intelligence" referred to the ability of a machine to mimic human intelligence to a point where it is indistinguishably sentient (Anyoha, 2017). In a contemporary sense however, the term has come to describe a variety of types of algorithms that allow a computer to learn patterns in large amounts of data and make predictions when presented with a new problem similar to its training. The high-level expert group on AI (AI HLEG) of the European Commission defines AI as: "*Systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals.*" (European Commission 2019). We agree with this definition and add to it that these systems also have a black-boxed dimension, where the algorithm is incomprehensible or undecipherable due to either accumulated complexity or self learning. Although not sentient, AI algorithms allow the computer to solve tasks on its own that are usually reserved to humans or require extensive or impractical amounts of manual code to perform computationally (Agrawal et al., 2022).

2.1.1 The AI Effect

McCorduck observed in her 2004 book 'Machines who think' that whenever a successful AI application was developed the definition of AI changed. Thus AI researchers were a branch of computer scientists dealing with the problems computers could not solve yet. This has come to be known as the 'AI Effect' where problems that were previously considered AI problems were degraded to 'computational problems' once they were solved (Sheikh et al., 2023). Interestingly though, in newer times the term AI has seemed to stick as a general descriptor of a broad category of learning algorithms even though they exist and perform tasks in the real world.

2.1.2 Types of AI

Two of the major categories of AI are machine learning (ML) and deep learning (DL). DL is a type of ML and both are modeled after the structures of the human brain in so called neural networks. Neural networks consist of layers of nodes. There are three types of layers – input, hidden, and output. There can be multiple hidden layers, and generally ML algorithms have three or fewer layers, and DL have more than three and usually hundreds of layers. Layers consist of neurons or 'nodes', which are connected to nodes in the next layer (IBM, n.d.-a). Each node on the first level of an image recognition model could for example correspond to an individual pixel on a digital picture.

ML comes in three flavors – supervised learning, unsupervised learning and reinforcement learning. A supervised ML algorithm is usually fed a corpus of training data which has been labeled and annotated by humans. A classic example of this type of ML is the ability to correctly categorize pictures of animals it has never seen before based on features it learned to recognize from the input data (IBM, n.d.-b). Other examples include email spam filters, predictive analyses of business decisions based on past experiences and more.

An unsupervised ML algorithm is fed a corpus of unlabelled data with the objective of finding patterns or structures that maybe aren't obvious to a human (IBM, n.d.-c). This usually yields a sorting of a dataset into clusters. This could be clusters of user habits revealing new archetypes for better targeting of advertisements or products, or detection of anomalies in large amounts of data.

Reinforcement learning is an interesting approach to machine learning reminiscent of animal psychology. An agent is 'let loose' in an environment authentic to its intended purpose. The actions of the agent are interpreted by a reward system and if it did well it is used as the basis for the next generation agent(s). The agent aims to repeat the behavior that rewarded it, but is also programmed to explore new actions. Depending on processing power thousands of these agents can be simulated simultaneously. The objective could be an obstacle course and the agent a virtual race car. The unit that gets furthest is used as the basis for the entire next generation of race cars. Once all cars complete the course, it is placed in a new obstacle course and the cycle continues until the algorithm has produced a race car that can master any course. A human can also be the interpreter of an agent's performance. Reinforcement learning from human feedback (RLHF) can be employed to fine-tune supervised or unsupervised ML algorithms (Bergmann, 2023). RLHF is for example used by the Icelandic Government for optimizing ChatGPT's ability to speak Icelandic in an effort to preserve the language (OpenAI, n.d.).

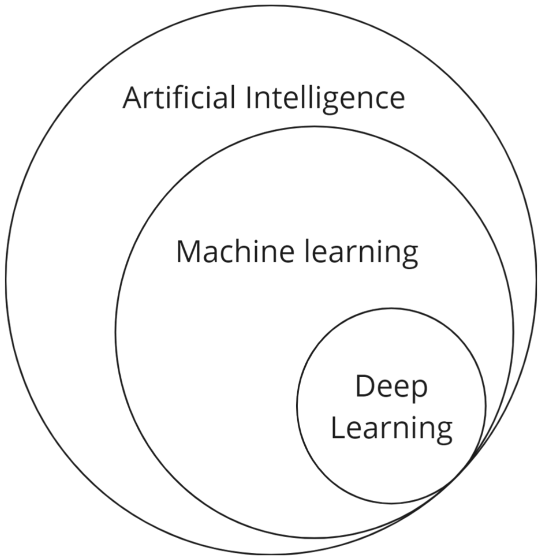


Figure 1: Relationship between Artificial Intelligence

2.1.3 User Interface

The user interface (UI) can be seen as the access point where users interact with designs like AI. The UI is the point of access that dictates what is available for use and constraints or allows for the different kinds of functionalities of a device. If an AI utilizes computer vision with machine learning and neural networks to derive meaningful interpretations of images, such as anatomical landmarks in ultrasound images, the UI is the point of access to such technology. The UI connects the inner workings of the AI with actionable information that the user can engage with.

The experience of those who have to use a system like an AI for ultrasound, is also influenced by who they are. If the AI fails to live up to the expectations of medical professionals and the working conditions of the clinic, it might be disregarded even if the AI is technically effective. The user interface can consequently result in user experiences that might lead to acceptance or rejection of the utility of the AI by the end user (Lieberman, 2009). The UI then plays a significant role in defining how the AI is actually used and perceived (Meske & Bunde, 2023; Lieberman, 2009). This is why it is important to study the users – not only how it should look and what elements or functionalities of the AI is available to the user through the interface – but also how they engage with it. This entails looking into the characteristics of the users and the task they have to perform with the AI. This includes how the UI portrays the delivery of information to the users and how they use that information in a clinical setting. The development of the UI is then a critical task in the design of an AI-system.

2.2 The Medical Field and Procedure

We work within the medical field of obstetrics and gynecology that is concerned with the care of pregnancy and childbirth, as well as treatments regarding women's health issues (Salomon et al., 2019). While gynecology encompasses the prevention, diagnostics, and treatment of symptoms related to female reproductive systems, obstetrics is especially involved with the diagnostics and treatment during pregnancy, childbirth, prenatal care as well as the postpartum period (Sundhedsstyrelsen, 2023). A clinical tool often used in the realm of obstetrics is the ultrasound machines. Ultrasound is high-frequency sound waves emitted by a hand-held probe/transducer, that is directed at the internal bodily structures. These soundwaves are then reflected and translated into images that can be interpreted and used in medicine practices like diagnostics. This "window" into the body is ideal for examining something like a womb and allows for early detections of complications during pregnancy by enabling the clinicians to identify fetal abnormalities (NIH-NIBIB, 2023).
Every pregnant resident is offered two routine ultrasound examinations in Denmark. The 1. scan is offered at the 1. trimester in week 11 + 0 that can help to identify chromosomal defects, while the 2. scan offered at week 18- 20 can help to identify birth defects (NIH-NIBIB, 2023). Later in the pregnancy – usually around mid-

trimester – ultrasound examinations can be used to estimate fetal weight (EFW) for pregnancies at risk (Salomon et al., 2019). Screening for fetal growth abnormalities are an essential part of providing adequate prenatal and maternal care and to prevent adverse outcomes related to childbirth like macrosomia (Salomon et al., 2019). It is also a very common examination locally in Denmark as well as globally. Following the ISUOG Practice Guidelines (Salomon et al., 2019) to perform the growth scans to EFW at least three fetal biometric parameters or standard planes are needed: Head circumference, abdominal circumference and the femur diaphysis length. These biometric parameters are then measured using biparietal diameter and various formulae. The calculations enable the clinicians to approximate a birth weight needed for diagnosis of fetal growth disorder and the planning of timely interventions (Salomon et al., 2019). These growth scans are often done by sonographers as operators, but can also be done by specialized nurses, obstetricians or by young doctors in training. This examination can however be difficult and requires many hours of training to become proficient in this procedure (Andreasen et al., 2021). The operator is accompanied by the patient in question, guided by clinical guidelines, their training and sometimes supervised by a senior clinician. This image-based examination and the quality of it, is dependent on the skill of the operator. They conduct the scan with the main purpose of detecting any growth anomalies, by collecting evidence, and deciding upon clinical decisions. This is done while manipulating the ultrasound probe and adjusting to ultrasound images provided by the ultrasound machine. When the operator is confident enough with the collected evidence used for interpretation, the examination is then concluded (Andreasen et al., 2021). It is within this clinical space that the AI has to be operational by supporting the clinician, underlining their clinical decision making. The AI in development is not supposed to substitute the operator by predicting diagnostics, it rather augments the skills or capabilities of the clinician by supporting their decision making and diagnostics. This emphasis on human-AI collaboration is a central part of AI-development. When performing a fetal growth assessment with ultrasound, the AI can help the operator to assess the picture quality and identify the biometric parameters needed to conclude the examination. Hereby addressing the uncertainty that the clinicians may face when performing scans for EFW.

2.3 Pioneer Center for AI

In the following sections we will provide an explanation of our external collaborators. Who they are and what we have been working with during this thesis.

On top of the hill in the botanical gardens in Copenhagen is an old observatory. Although not made of ivory, this tower is supposed to house the very top of Danish academia working with AI. The Danish Government has funded the Copenhagen Pioneer Center for Artificial Intelligence (PCAI) with 354 million DKK over 13 years (until 2034) and more centers are being planned in other major Danish research cities. The PCAI is partnered with the University of Copenhagen (UCPH), Aalborg University (AAU), University of Aarhus (AU), the Technical University of Denmark (DTU), and the IT University of Copenhagen (ITU). It is through these universities that most of the funding is distributed as grants for research, postdoc and ph.d. positions (www.aicentre.dk/the-centre-p1).

2.4 SONAI

One of the established research projects at PCAI is the SONAI-project that consists of an interdisciplinary team of professors, external collaborators, scientists, Ph.D's and a few master students affiliated with a variety of different other institutions.

Like many other PCAI projects, the technical team consists mostly of Ph.D. students specializing in relevant algorithms led by professors from DTU and UCPH. The head of the research project is a chief physician in obstetrics from Rigshospitalet, UCPH and CAMES. The project is divided into several work packages that each has their own raison d'être. Every second Thursday everyone affiliated with the project is invited to the old observatory to discuss, collaborate or present findings. Our collaboration in this research project is within the space of the clinicians and the engineers. Within the SONAI is several parallel projects ranging from AI heart anomaly detection to early birth detection.

2.4.1 Progressive Concept Bottleneck Model

Since the technical details of how the AI is built is beyond the scope of this thesis, we will only outline the setup of the AI and explain some of the key terms. It is this installation of the AI that we have employed in our data collection. The source is based on first hand experience and unpublished papers from some of our collaborators at the SONAI-project. The current AI-model is trained on a dataset with screened images and pregnancy outcomes from a large fetal medicine database containing over half a million pregnancies over the last 14 years. Furthermore the training includes a large dataset with manually annotated anatomical structures and image quality features. The model behind the AI is a Progressive Concept Bottleneck Model (PCBM) that is a machine learning model. The choice for the PCBM is centered around comprehensibility in clinical decision-making processes, by providing predictions that are interpretable by humans. In this case the predictions align with the clinicians' natural thinking process. An example could be as illustrated in Fig. 2. First the clinician identifies an anatomical landmark like the femur by looking at the ultrasound images. If the images contains the property of what constitutes a standard plane eg. the correct angle of the femur or that both ends of the femur is visible, then it must be concluded as a standard plane.

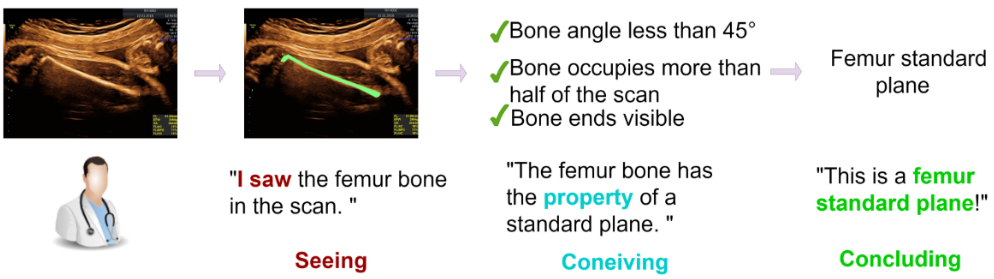


Figure 2: The concepts of "Seeing", "Conceiving" and "Concluding" as a clinician as well as in the PCBM (Lin et al., 2022)

The AI functions in the same way by utilizing a network that learns to identify the properties of an image known as conceiving. A predictor model is then applied that concludes from these properties. Here the predictor learns from annotations and from training. When the prediction is concluded, visual cues are then presented as explanations in real-time while performing ultrasound. When performing an EFW, the clinician is then confirmed in their clinical decision making with AI-feedback that helps to conclude whether the images are in a standard plane or not. As seen in

Fig. 2 the AI is capable of identifying these relevant anatomical landmarks like the Thalamus, or features that indicate a wrong plane like the Fossa Posterior, while it is also capable of assessing the overall quality of the image. If the anatomical landmarks are identified and the quality of the image is satisfactory, it is then possible to place the "Calipers" needed for measurements.

This functionality to conclude on planes is the transformative capabilities of the AI and is the outset for the features to come. One of these novel features in the making, is a guidance system that can aid with suggestions on how the clinician can reach the next standard plane needed to EFW. This feature could guide clinicians, especially novice clinicians, into the correct standard plan and be utilized as a learning tool or bridge the gap in areas where ultrasound expertise is missing.

2.4.2 AI-Deployment in the Clinic

One thing is the development of the AI itself, with all the training data and the AI-model. However, bringing the AI to real-world deployment in clinical workflows can be very challenging. The success of the AI depends to a high degree on how useful it is in a clinical setting, where it ultimately has to function. In the SONAI-project the interdisciplinary collaboration between engineers and clinicians has allowed for early clinical testing. By bringing the AI to the clinic in its early stages it is possible to examine early failures and rectify them early in the development phase. One of the challenges was how to bring the AI to the clinic for testing and what kind of setup is feasible.

As seen in Fig. 3 the AI does not stand alone, but needs a whole setup to be operational. It might look "simple" at glance but the whole framework leverages several technologies to capture, process and present real-time video feedback from the ultrasound machine.

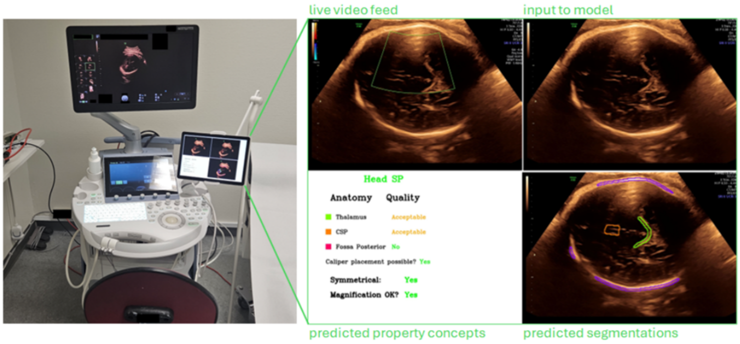


Figure 3: The AI Setup in the clinic

An illustrative explanation of the system architecture can be seen at Fig. 3. This installation has allowed for the AI model to be deployed in a real-world or simulated clinical setting, where the operator receives immediate AI-feedback during the ultrasound scans. Following this architecture the Ultrasound machine is connected to a portable mini PC (NUC) with a Server, via a HDMI-to-USB converter. Within the server are the Docker containers that perform various computational tasks that constitute the AI-system. The output can be accessed via a simple web application through a tablet that is connected to a closed local network. The feedback is then immediately visible on the tablet next to the ultrasound machine as seen on Fig. 4.

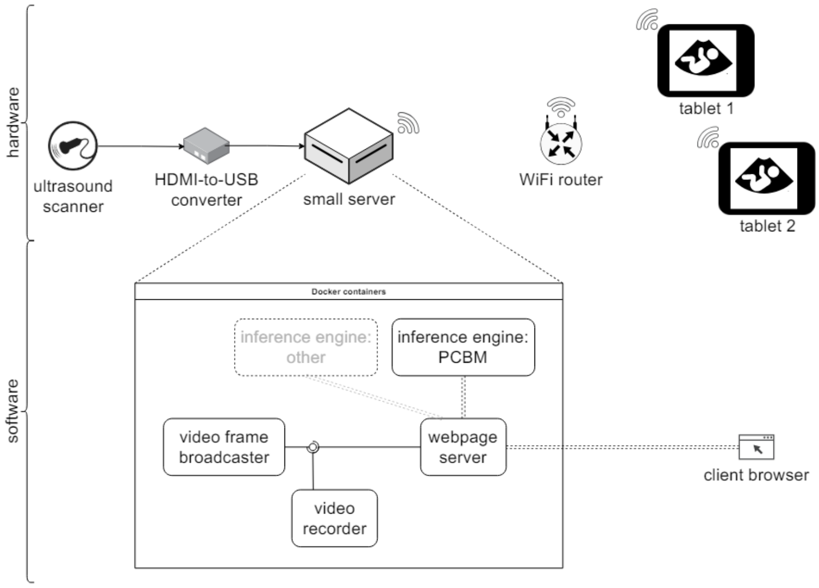


Figure 4: System Architecture Illustration (*Unpublished Article, SONAI*)

3 Theory

In the following sections we will present our theoretical framework and detail some of the concepts that are also presented in our article. These will therefore be expanded upon here with additional threads being drawn to relevant social scientific and philosophical literature. First we will present Value Sensitive Design (VSD) as a discipline for user-oriented responsible innovation and design with inclusion of some newer approaches to design issues such as AI. Then we will present the philosophical framework that informs both our value frameworks and the VSD methodology we employ.

3.1 Value Sensitive Design

Value Sensitive Design (VSD) is an approach used in interdisciplinary design with the purpose of developing design requirements that will enhance a technology's compatibility with (a) society. VSD contrasts a 'traditional' view on technology as being value-neutral artifacts developed to accommodate some functional requirements set by a user or customer (van den Hoven et al., 2015). Seeing technology as neutral and user values as externalities has become increasingly difficult as it plays a larger part in our daily lives, with the advent of digital and information technologies, and it is in the wake of these technologies the first VSD frameworks were developed (for example Friedman & Kahn, 1992; Friedman Nissenbaum, 1996). VSD is a normative discipline in the sense that it deals with how a technology 'ought to be' designed to fit in a specific context in addition to the hard technical requirements.

3.1.1 Central Concepts

Batya Friedman and Peter Kahn of Washington University are considered the founders of VSD as a design discipline. Friedman has produced a body of foundational literature exploring both the methodological approaches as well as the philosophical and theoretical aspects of VSD. Values in VSD are defined by Friedman as "*what is important to people in their lives, with a focus on ethics and morality*" (Friedman & Hendry, 2019, p. 24). This human-centered approach to values is also recognizable in van de Poel and Royakkers' definition of values as "... *lasting convictions or mat-*

ters that people feel should be strived for in general ..." (van de Poel, 2013, p. 27). Friedman operates from a set list of values, usually human welfare, ownership and property, privacy, freedom from bias, universal usability, trust, autonomy, informed consent, accountability, courtesy, identity, calmness, and environmental sustainability (Friedman & Hendry, 2019). Other VSD authors have criticized using a set value list as it unavoidably benefits specific stakeholder groups over others, risking exclusion and reinforcing current privileges (Borning & Muller, 2012). Although Friedman acknowledges the obvious shortcomings of a heuristic list of specific values she ultimately deems it more productive as it allows her and other scholars to build upon previous work (Friedman & Hendry, 2019).
van de Poel (2013) describes 'norms in VSD' as design principles to operationalize or achieve the values:

> "*One kind of norms that are especially important in design are end-norms* (...) *because design is aimed at the creation of technical artifacts or at least blueprints for them. [They] may refer to properties, attributes or capabilities that the designed artifact should possess*"
> (van de Poel, 2013, p. 19).

Thus, design norms can be described as the features of a technology, specified in a non-technical language, that it needs to have in order to realize the values. We see VSD as the translation of abstract values of a society and smaller groups of people into technical design requirements.

3.1.2 Tripartite Methodology

The traditional VSD approach employs a 'tripartite methodology' (Friedman et al., 2003). Following this methodology, the design team will be put in a position where they are able to address some of the value implications of the technological design. Values shape technology development. Without proper inclusion of the values of the actual users or indirect stakeholders, the technology will be imbued with the values of the designers and product owners. This may not even be to the designers' own best interests. In this sense, VSD is both a learning process and an unlearning process, as the design team uncovers where their own values may have influenced conceptual design decisions, so that they can adjust accordingly.

The first step of the tripartite methodology is the conduction of the *conceptual investigations*. During these, the design team addresses the different frameworks and boundaries: Who are the direct and indirect human or nonhuman stakeholders? Which values do we anticipate coming into play, and how might we order these? Which philosophical and theoretical frameworks do we use to support our methods and decisions? This step is followed by the *empirical investigations* where the human context in which the technology is expected to exist in is explored. This can be achieved utilizing elements in the entire social scientific toolbox. If a heuristic list of values is used, in which order would the stakeholders list these themselves? If values are sourced directly from the stakeholders, how are these uncovered and ranked?

The last step is the *technical investigations*, where potential and existing technologies are investigated both retrospectively and proactively, followed by the design of a novel product (Friedman & Hendry, 2019). These three phases can be carried out in series, simultaneously, in a different order and iteratively.

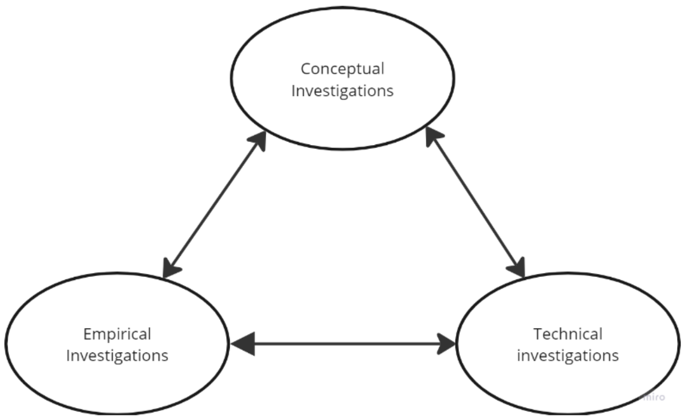


Figure 5: Model of the Tripartite Methodology

Other VSD authors' adapted methodologies usually relate to these original three phases. For example Umbrello & van de Poel (2021) relate their adapted four-phase approach to responsible AI-design directly to the tripartite methodology in their model. (Fig. 6).

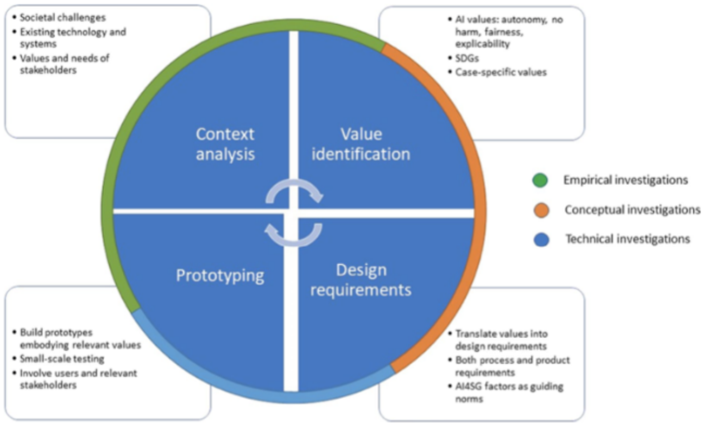


Figure 6: The Model of the Tripartite Methodology by Umbrello & van de Poel (2021)

The original tripartite methodology is also present in the first iteration of the "Values that Matter" approach to VSD (Smits et al., 2019) that will be introduced later in this chapter, especially in the first two phases "explore" and "conceptualise" (Fig. 7).

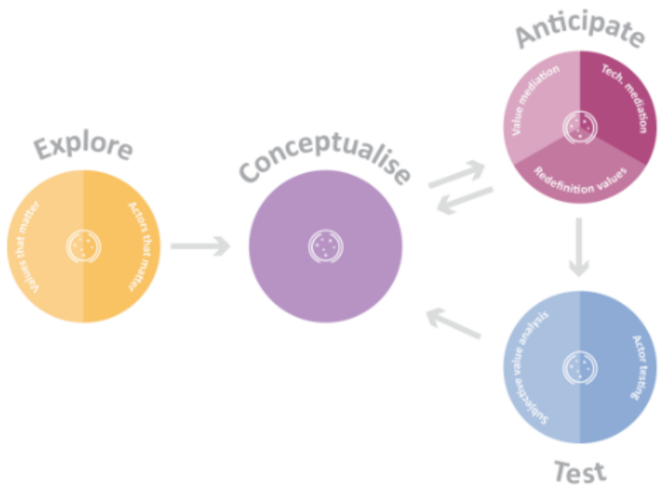


Figure 7: The Model of the Tripartite in VSD by Smits et al. (2019)

All this to say that the tripartite methodology is at the core of VSD either explicitly or implicitly, and the specific methods are not set in stone. As a researcher you are expected to understand relevant literature of conceptual ethical and theoretical discussions, perform investigations with stakeholders and translate this to technical specifications.

3.1.3 VSD in Artificial Intelligence

Steven Umbrello and Ibo van de Poel (2021) responsible design of AI technologies. They argue that the self-learning aspects of ML algorithms pose a novel challenge. Therefore even VSD approaches proven useful for designing digital and information technologies do not suffice for ethical design of AI. They describe their modifications to the VSD-framework as follows:
“The modifications we propose are threefold: (1) integrating AI4SG principles into VSD as design norms from which more specific design requirements can be derived; (2) distinguishing between values promoted by design and values respected by design to ensure the resulting outcome does not simply avoid harm but also contributes to doing good, and (3) extending the VSD process to encompass the whole life cycle of an AI technology to be able to monitor unintended value consequences and redesign the technology as needed.” (Umbrello & van de Poel, 2021, p. 288).
For this study, we only adapt the second modification, i.e. the ‘value identification’-stage from their methodology (also visible on the model shown in the former section 3.1.2), as the other modifications are not as relevant for our UI-focus.
Umbrello and van de Poel’s mixed approach of value identification is both informed by normative dimensions and derived from the specific context. Some sort of value list for typical AI issues can prove productive, but it is also important to elicit values from the bottom and up (Umbrello & van de Poel, 2021). They thus provide three sources for human values:

- **Values promoted by the design**: These values are an orientation towards contributing to the betterment of society, such as contributing to the UN’s Sustainable Development Goals (SDG’s).

  “*For a VSD approach to AI to be more than just avoiding harm and actually contributing to social good, there must be an explicit orientation toward socially desirable ends.*” (Umbrello & van de Poel, 2021, p. 288)

- **Values respected by the design**: Values can be respected by design through intentional integration of specific ethical principles with clear guidelines and definitions on what is good or potentially damaging for e.g. all humans. Such guidelines could come from religion, human rights etc. The authors source these from the EU high-level expert group on AI (AI HLEG) European Commission 2019, but could also be another moral framework with a societal lens. We agree with and build upon the AI HLEG's principles of respect for human autonomy, prevention of harm, fairness, and explicability. These will be explored more in depth in the analytical chapter.

- **Context specific values**: these are the values that are identified in the context and not covered by the other two sources.

Umbrello and van de Poel do not specify which type of AI developers they intend to adopt this moral design-framework for, and we can only speculate about how realistic it is to be used for AI-development within profit-driven private companies. It does however fit well within research-driven public projects such as the SONAI-project, and with the unique insights of Danish public companies into databases associated with the CPR-registry, a simultaneous focus on not only preventing societal harm but increasing social good can prove to be a promising approach.

3.2 Technological Mediation

While Umbrello and van de Poel's framework provide a robust VSD approach to the design of responsible AI, and while much of their concepts apply to our investigation, we still need another addition. The authors have not provided a method for eliciting values from the bottom-up (Context specific values) and with our UI-focus and specific user-group, we chose to integrate this approach with another VSD-methodology, namely the *values that matter* (VtM)-approach which operates strictly at a user-level. The VtM approach is based on *technological mediation* which is a theory within the philosophical branch of post-phenomenology.

3.2.1 Post-Phenomenology

Phenomenology as a philosophical branch is preoccupied with experience. The objective of a phenomenological study is to uncover the structure of experience which is usually done in the offset of a human subject in relation to objects in their surroundings (Armstrong, 2005). The notion of subjects interacting with and in the world is expanded upon by the American philosopher Don Ihde and later the Dutch philosopher Peter-Paul Verbeek who claim that technology is the key to understanding human experience. Contrary to earlier phenomenologists, to Ihde and Verbeek technology is not just objects in the world that we human subjects can interact with, nor is the world not just an abstract collection of raw materials waiting to be processed into technology (Radboud Reflects, 2018). Post-phenomenology posits that technology isn't merely external to humans, but something that mediates the world, and changes our way of interacting with it and each other. New digital technologies didn't just raise privacy questions but have fundamentally redefined the concept of privacy itself from something spacious and tangible to a question of data ownership (Kudina & Verbeek, 2019). The experience of subjects in the world is radically transformed when mediated by technology.

When seeing the world through a pair of glasses or the moon through a telescope, the technology becomes an extension of the body. Our arms embody and act the intentions of our brains, and holding a tool, driving a car etc. merely extends our body further into the material objects. Ihde (1990) calls this an embodiment relation, described as *(I-technology)* $\rightarrow$ *world*. This description illustrates how the human does not merely use a technology to perform their intention in the world, but the human-technology hybrid does. A person who has a severe visual impairment and functions as a university hospital surgeon and doubles as a professor and researcher relies totally on their glasses. However, they are nearly invisible and taken for granted – no one thinks the credit of the achievements of the surgeon ought to be shared with the glasses. Though they are nearly indistinguishable, it is still through the (*surgeon-glasses*) combination that the surgeon performs their praxis in the world.

Conversely, technology can also enter into a near indistinguishable relation with the world, described as *I* $\rightarrow$ *(technology-world)*. This is the hermeneutic relation. The same surgeon might look at instruments to determine the state of the patient. It is not possible or practical to determine the blood pressure and heart rate without an instrument providing the numbers. Thus the world or the phenomenon is hidden behind a screen with numbers that the surgeon reads and determines a decision upon.

3.2.2 Mediation of Values in Technology

A central concern in (social) studies of technology is the 'dilemma of control' (Collingridge, 1980), which has come to be referred to as 'the Collingridge dilemma'. If a technology is early in development, plenty of the design decisions can be changed, but the design team does not have sufficient knowledge about how it will affect society. Conversely, if the technology is already on the market, plenty of actionable analyses start to emerge as its interplay with social life becomes observable, but then it is often too late to change the course. Thus for researchers occupied with responsible design of technology – not least VSD researchers – the Collingridge dilemma remains one of the biggest challenges (Kudina & Verbeek, 2019).

Technological mediation offers a nuanced approach to addressing the Collingridge dilemma when it comes to values. In relation to values, Kudina and Verbeek call the relationship between technologies and values "Technological Value Dynamism". As an example, they investigate the technology of Google Glass, a sort of smart glasses with a screen and a camera. Google released an "explorer" version of the technology to a limited number of people to test the device. Among these were tech influencers on YouTube who posted their experiences to the platform. This allowed the researchers to study people's reactions to the glasses, especially in the comment sections. Since the glasses as a product were at the threshold of society – that is developed as a prototype and used in authentic situations but not fully released – this allowed Kudina and Verbeek access to understanding the mediation of the value of 'privacy' by these glasses.

Privacy means something different in real life compared to online. An example of the online definition of the value of privacy can be seen in the regulatory steps made to protect it such as the EU's General Data Protection Regulation (GDPR). GDPR specifies that consent must be given for any user data to be processed in any capacity by the owner of a website. Privacy means data ownership. In the real world however, privacy has a spatial and interpersonal dimension. It's tangible, if people are far away they can't hear you, and if you don't know them, they probably won't remember seeing you. Even in public spaces, studies have shown that people have a reasonable expectation of privacy (e.g Roessler & Mokrosinska, 2013). There is a reasonable expectation of the indifference of other people and that they will forget you. Google glasses with their potentially always-filming cameras introduces the online "data-ownership" definition of privacy to the real world. The prospect of people seeing you can imply that they would remember you looking back at their footage. The value of privacy in public could change.

This investigation of attitudes and reactions on a technology at the cusp of society is not for the purpose of understanding how to make that technology better in a traditional sense. It will help understand the compatibility with societal values.

3.2.3 Mediation in VSD

Technological mediation has been methodized into an adaptation of VSD, namely the Values that Matter (VtM) approach (Smits et al., 2019, 2022). By using technological mediation VtM provides a framework for understanding how values themselves are transformed by new technologies. This can be used in the process of responsible design of technology, by allowing designers to anticipate technologically induced value mediation in an informed way.

The authors operationalize this by introducing their own tripartite methodology, focused on finding the "right balance between speculation and experimentation" (Smits et al., 2022, p. 43).

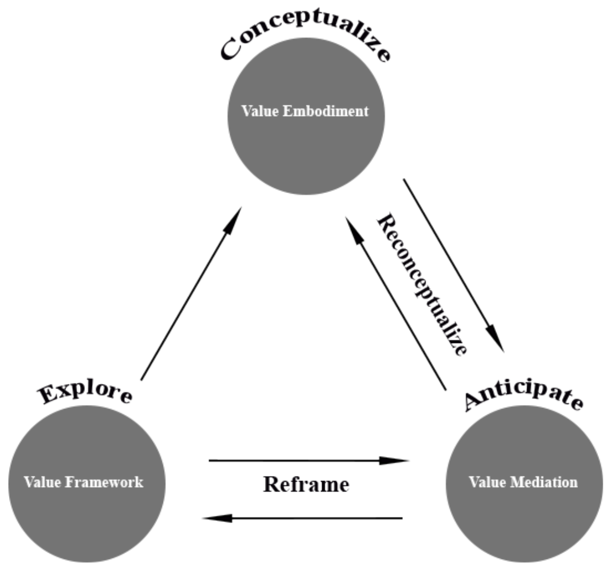


Figure 8: The Values that Matter Model (Smits et al., 2022)

This is the latest iteration of the VtM model. We showed an earlier iteration in 3.1.2 which was a bit closer to the tripartite methodology of VSD. Although methodologically adjusted, the three phases of VtM are based on the conceptual, empirical, and technological investigations of VSD (Smits et al., 2022).

- **Explore** – This phase involves "*mapping out the context of the design problem*" (Smits et al., 2019, p. 399), which includes identifying important actors and developing a value framework. This phase combines conceptual and empirical investigations.

- **Conceptualizing** – During this phase, a prototype is created based on the defined groups and values, incorporating the technical investigations. "Conceptualize" in this case refers not (only) to building a philosophical and theoretical framework for the continued investigation, but more importantly to developing a technical framework – a concept. This could be a drawing, 3D model or a working prototype. When the conceptualization phase is revisited after exposing the concept to the actors, then it is a case of re-conceptualization.

- **Anticipation** – This phase operationalizes and stresses the importance of measuring the effects of technology on the meanings of the values themselves.

According to the authors, technology fundamentally redefines how humans engage with values. The anticipation phase is a way to strike a balance between speculation and experimentation. This can be done in practice by introducing the concept to the relevant actors in an authentic context. Thus it can be examined how a technology mediates the identified values differently than in a context without the technology (or with a previous iteration of the technology). These value experiences can be identified and translated into design requirements and make the basis for a re-conceptualization.

3.3 On Stakeholders and Society

Different VSD authors have different approaches to the breadth and depths of identifying and designing for the stakeholders. Stakeholders can be anything from individuals, to groups to entire societies, animals and past and future generations (Friedman & Hendry, 2019). Traditionally, VSD has distinguished between direct stakeholders and indirect stakeholders. The direct stakeholders are those that the technology will interact with as users, and the indirect stakeholders will be affected by it, but not interact with it. The clinicians are the direct stakeholders of the AI we are investigating, as they are the direct users, and the UI is designed to make their work easier. The patient receives the benefit or harm with the system, but is not interacting with it directly. The fetus is likewise turned into a patient by this technology (Verbeek, 2008), its vitals being investigated and decisions with outcomes that can affect its life outside the womb are being made. Society as a whole is also an indirect stakeholder due to the economics of development and integration of health care services.

We employ two different VSD approaches, the VtM-approach and the AI specific VSD approach. We see VtM as a techno-anthropologically compatible adjustment of the tripartite methodology of VSD specifically due to its solution to the Collingridge dilemma and its emphasis on obtaining a deep understanding of the direct stakeholder's values. We do however find it difficult to apply this approach to an investigation of the indirect stakeholders, and in the context of AI, we find it unable to include societal values into the framework. Conversely Umbrello & van de Poel

(2021)'s AI specific VSD approach goes very deeply into societal values with a focus on indirect stakeholders and future generations. We find its suggestion to sourcing 'context specific values' vague and too detached from user groups or direct stakeholders. Our approach seeks to use VtM as the primary methodological framework, and Umbrello & van de Poel (2021)'s three sources of values as a secondary framework that allows us to include broader societal values specific to AI in the design requirements.

4 Methods

In this chapter we will explain the method of data collection that we have applied during this thesis. This entails how we secured access to the SONAI-project and external collaborators and how we applied qualitative methods with user-testing and interview to enlighten our research question.

4.1 Access to the Field

To start data collection we needed access to an increasing number of elements, everything from participants to hardware and software. We pitched the theme of our thesis to our external collaborators at CAMES that included us in the SONAI-project. The collaboration between us and SONAI has the nature of an associated relationship. That means that we are not integrated into the main workflow of the research grant and do not have any formal obligations to the organization or funding bodies.

Fortunately we could join the engineers as well the clinicians with aligning interest in a collaborative environment. We were then granted access to everything we needed for data collection as well as technical assistance and clinicians as participants. One request from the SONAI-project was that the setup should function as a preliminary setup for future research endeavors. This collaboration was then lucrative in the sense that we could assist each other and gain important insights. Among these was insights into how to make the first iteration of the AI operational as well as a first look into what dynamics are at play when clinicians engage with AI in clinical decision making.

As mentioned in Section 2.4.2 we had access to Docker with the AI-elements through remote access via Visual Studio code. This allowed us to run code with a NUC-mini PC with access to a server to make the AI operational. We acquired all the equipment illustrated at Fig. 4 including various ultrasound machines from CAMES and Rigshospitalet, all the hardware, as well as advanced ultrasound phantom models of a pregnant patient used for simulation. Fig 9 illustrates the first iterations of the whole setup that allowed us to conduct the first testings in the Engineering garage at CAMES and later at Rigshospitalet.

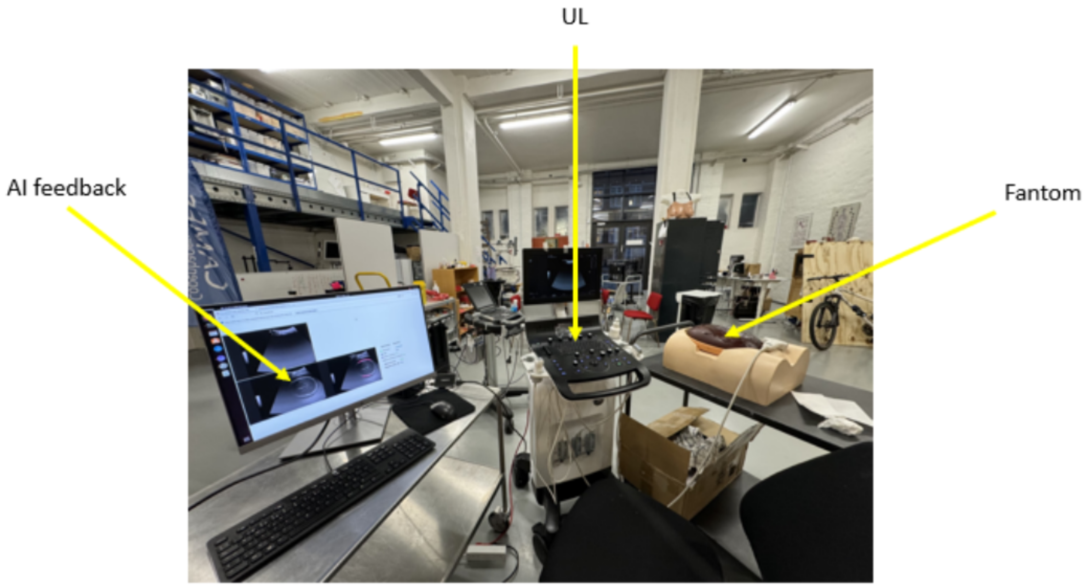


Figure 9: The AI-setup in the Engineering Garage at CAMES

4.2 Ethnographic Reflections

This master's project is a continuation of last semester's internship of Andreas von Essen (ABE) at Copenhagen Academy for Medical Simulation (CAMES). During this internship period and as a method for the internship report, he employed ethnographic methods including participant observations. This includes several visits to the obstetrics department at Rigshospitalet, where they perform the ultrasound procedures, as well as collaborative meetings at PCAI, CAMES and SONAI. The internship provided the foundation for this thesis, which includes networking, arranging permissions and other preliminary preparations as well as insights into working conditions of the team. We mention this because this prior relationship is interesting in relation to the anthropological concept of 'gatekeepers' (e.g. Hammersley Atkinson, 2019). A gatekeeper is a person who holds an authority to allow or deny a qualitative researcher access to other people (Roller & Lavrakas, 2015). This can both be a formal gatekeeper following rules or bureaucracies of access or an informal one who needs to be wooed. In our case, there is an actual formal process of becoming an associated master's student of the PCAI. You need to send an application to become affiliated at *www.aicentre.dk/affiliation*, then if accepted, you gain access to a portal where you can respond to calls for student projects. As ABE met regularly with the SONAI-group during his internship (as CAMES, PCAI and the SONAI-group have crossover members) this opened up an informal channel for gaining affiliation outside the standard procedure.

Although the SONAI project consists of a group of people with a somewhat shared set of values and practices, investigating them as a cultural group ethnographically has not been a central part of this thesis and made little sense to our project. Not only are they rarely to find at one place at one time – short for the collective touch-base meetings on Thursdays – we are not doing a meta-investigation of our collaborative partners. That said, it is important to understand how they work to get an idea of their expectations of us, the ambitions of the project and how they interact with and perform the technology in question. More importantly though, ethnographic methods came into play when we performed the user testing with the potential users of the AI alongside members of the SONAI-group. By being a part of the user testing team it was possible to witness these interactions firsthand and to have a role as designers. We were able to examine the dynamics at play when introducing AI as a supportive tool in the clinic, but also to examine how to make it operational. This included the setup and the elements of the user interface. Ethnographic methods facilitated the discovery of implicit knowledge and unarticulated needs among users. Although an initial disadvantage for Andreas Pedersen (ASMP) joining the project in February in the middle of the planning as more of an outsider, it provided a natural process of stepping back, reflecting on the tasks and interactions in ways that revealed deeper insights into the expectations and preferences of the users.

4.3 Interviews

Interviews were our primary method of empirical data collection. We employed the semi-structured interview approach, which is the most used type of interview in modern research (Tanggaard & Brinkmann, 2020). The semi-structured interview is a blend of pre-formulated questions with ample room to diverge into unplanned naturally occurring themes during the conversation. Interviews grant access to the life world of the participants – their perceived experiences with phenomenae. However, it is important to note that being told about experiences will never let the interviewer really experience that which the interviewee is explaining. Moreover, there is also a discrepancy between what people say they do, and what they actually do. We drew inspiration from James Spradley (1979) when preparing our questions. Spradley's interview approach is ethnographic, meaning that the interview becomes not about getting answers for your preconceived questions. To Spradley, discovering the right questions is the key to understanding the interviewee.

This approach is particularly relevant in our study, where using technical design terms to ask about screen elements and UI placements would be ineffective. Clinicians and technicians operate within different cultural frameworks, influencing their perspectives on workflows and decision-making. Spradley highlights that traditional interviews often fail because questions and answers stem from different cultural meaning systems especially if the aim is to be able to emphasize with the interviewee and understand their social or societal situation more intimately. Therefore, this "ethnographic interview" views the question and the answer as the same element in human thinking – questions imply answers and both must be discovered from the participants before you can begin to claim to understand their life world. By adopting this approach, we did not only aim to bridge the cultural gap between clinicians and technical designers by ensuring our questions resonated with the participants – we wanted to be able to put ourselves in the shoes of the user when designing our UI-mock-ups.

An example of us asking a question from a different cultural world than the clinician is during our interview with a participant (referred to as P5). We asked about their trust in the system.

Excerpt from transcript Interviewer: "*Would you trust the AI to be used in the clinic?*"

Participant: "*Yes, 100% I am going to trust the AI-interpretations. if it is approved for clinical use it must be good it enough*"

This question was based on the world of the AI HLEG and our own theoretical framework. But we learned that this question does not make sense in a clinical context. To a certain extent, a fundamental trust in a medical device is not something doctors would be concerned about, as its very presence in the department means that it must be good enough to be there. This early interview changed the way we approached the question of trust in the AI and the way we wrote our coming interview guides.

4.4 Focus Groups

Although focus groups can mistakenly be equated to interviews just with more people, the data it produces is different. They produce data about a specific social situation related to a subject of the researchers choosing, where interviews produce knowledge about an individual's life world, experiences and opinions. Both interviews and focus groups can be seen as expressions about social action albeit expressed in different contexts (Halkier, 2020). Normally, the selection of participants in a focus group is a methodological exercise in itself, but as it was, we were able to get access to a diverse group of relevant participants without any recruiting by gaining access to the morning meeting of the gynecological department at Slagelse Hospital. There were nurses, sonographers and doctors of different specializations, age and experience with ultrasound scans. A downside to this approach could be the fact that many of the participants of this focus group constellation would likely know each other beforehand, and whether or not this has a positive or negative effect on the outcome is a lively debate in the focus group literature (Halkier, 2020). In

our case, the workplace hierarchies in particular could play a role. A junior clinician in a focus group with their supervising doctor may be less inclined to air challenging opinions, whereas there may be a more free space for critical conversation had they not worked together. Alas, we did not notice this, and observed that the clinicians gave each other space to talk in a respectful manner.
The session was in three phases. First, the participants were given a short lecture about the project and the capabilities. Then the participants were shown pictures of the UI-mock-up and asked specific questions about the elements and features on screen. Lastly, there was an open conversation about both the AI and the UI. We did not moderate the conversation more towards UI topics, as we observed that the conversations about the use and functionalities of the AI naturally led to discussions about the UI and vice versa.

4.5 Human Centered Design

We employed a multifaceted qualitative research approach for data collection. The primary method of data collection was semi-structured interviews following an interview guide (see Appendix 2). This interview guide was employed in both presentations with AI mockups and during/after AI-usertesting. The guide was purposely designed to explore what dynamics are at play when introducing AI in a clinical setting. Furthermore the methods employed in this thesis follows a scientific protocol that was reviewed by researchers in the SONAI-project. To begin the data collection the protocol was sent to The Danish National Committee on Health Research Ethics as a request and approved (see Appendix 9).

4.5.1 Focus groups interviews with a presentation of AI/UI mockups.

The participants were presented with UI ranging from simple AI-feedback to more advanced interfaces containing elaborate descriptions with identification and translations of captured ultrasound images. The semi-structured interviews allowed for discussions about the benefits, challenges, and ethical considerations about the introduction of AI in a clinical setting. These interviews were also concerned with the norms and values in the clinic and in which way AI could affect those. This allowed the participants to share their perspectives on AI's role in fetal weight estimation in obstetric ultrasound and to articulate their impressions and perspectives while being presented with different UI's. It was then possible to identify common themes regarding AI, based on the UI presentations and the mediating impact of the technology. The UI mockups were made in CANVA and can be seen under Appendix 4, 5, 6, 7. The mockups were constructed to be in continuous development as the data collection was progressing. It was possible to iterate several designs exploring different UI elements. Sometimes the participants also had concrete design requirements to be considered. As the data collection was concluded the mock-ups were the basis of the final conceptualization.

4.5.2 Usertesting workshop with AI-prototypes.

The user testing simulated a clinical setting where AI was introduced as the participants performed ultrasound to do a EFW. As mentioned in Section 2.2 the fetal biometric parameters needed to conclude EFW includes: Biparietal diameter, head circumference, abdominal circumference and femur diaphysis length. The physical setup consisted of an AI-prototype running on a local computer that provided real-time predictions of the quality of the ultrasound image to the participant while performing EFW (see Fig. 4 and 9). Then the participants had to identify the three standard planes to conclude the usertesting. The participants were actively engaged throughout the workshop and were encouraged to vocalize their thoughts and observations out loud, providing their immediate insights and experiences while interacting with the AI. Since this can be very challenging due to cognitive load, there was also time for further discussion following the interview guide after user testing. During the user testing the participants were surrounded by at least one observer that also facilitated the setup and one interviewer. This approach led to a dynamic exchange of feedback and allowed us to capture the participants' first reactions and preferences as they interacted with the AI-prototype. All the interviews were recorded for later analysis.

4.5.3 Participants and sampling method

In our study, we utilized purposive sampling to select physicians specializing in gynecology and obstetrics as well as novice clinicians with little to no experience in ultrasound. This allowed us to have a controlled representation of individuals with a diverse expertise in ultrasound, offering a broad range of perspectives.

A total of n=22 participants attended our data collection. All participants had varying degrees of experience in performing EFW spanning from zero to many years of experience. The user testing was done in two settings at CAMES Engineering or at Rigshospitalet with respect to the clinician's time and demand in their corresponding clinical departments. The time available for each individual user testing with interviews was therefore varying.

4.6 Data Processing

Working with qualitative data – especially around smaller sample sizes – often poses the issue of how to protect the participants from potential repercussions that could come from voicing their opinions or even their participation in the study. As qualitative scientists we have a responsibility to not put individuals in a bad light, but we also have a responsibility not to censor the voices of our participants (Brinkmann, 2020). Although it is standard practice to anonymize research participants, removing every possible trace that theoretically could be followed back to them would prove unproductive. Combined with the perceived uncontroversial nature of the subject of this study, we assessed that pseudonymization would suffice – a literary face-blur and voice-scrambler swapping out obvious characteristics of the participants with fake ones.

We transcribed our interviews using a journalistic approach, meaning that we did not transcribe non-spoken elements such as mood, facial expressions or tone as is noted down in e.g. the Gail Jefferson system. For some audio files, we used the Good Tape website hosted by Zetland, based on the AI-model Whisper. This service is GDPR compliant, securely encrypted, instantly deleted after processing, not used for training, trusted by journalists, and Danish-based.

4.6.1 Coding

When coding qualitative data, there are two major strategies you can employ. One is a priori coding, and the other is in vivo coding (Swain, 2018). A priori coding is when most or all of the categories for coding are decided beforehand, usually based on a theoretical framework. In vivo coding is when the categories are derived directly from findings in the data. Our coding strategy was a mix of a priori coding and in vivo coding. From the VSD methodology, we knew that we would code for values. Therefore we decided that our top level codes would be a priori defined based on values. We then placed the design norms as child level codes. These categories were coded in vivo.

4.6.2 Data Ethics

In the role of qualitative researchers, we oriented ourselves towards Svend Brinkmann (2020)'s four rules of thumb for responsible ethical data collection.

1. **The first rule of thumb**: Informed Consent - This rule is twofold, to ensure that the participants know what the project is about, and to ensure that the participants agree to having their data processed by us. We made sure that all participants were informed that we were doing a study as part of our master's studies that built upon their expressions, and that we intended to publish an article based on the same. Participants were provided with our e-mail addresses or the contact info of the SONAI-group so they would be able to retract their consent at a later point.

2. **The second rule of thumb**: Confidentiality - In line with our principle of pseudonymization, we ensured that all identifying information would be distorted in our documents and appendixes, but only to the extent that the anonymization would not distort the data.

3. **The third rule of thumb**: Consequences - A big risk one takes when sharing a study with other people is that it might be read and influence the decision making around that area of study. Taking a responsible design approach, this study aims to influence the technical designers of the SONAI project to adjust the UI to be more compatible with the sustainable development goals as well as EU's ethics guidelines. On a higher level, this means that we contribute to the broader acceptance of using AI in and with medical devices as a whole and using Danish patient data for training AI.

4. **The fourth rule of thumb**: The role of the researcher - It is important to stay true to our actual role as master's students from AAU. We are not clinicians, AI engineers or even SONAI-researchers. Staying true to our role is important as to not directly or indirectly deceive anyone into sharing more than they need to.

5 Analysis

In this chapter we will detail how we have approached the analysis using VSD and VtM as a methodological and theoretical outset to unfold our empirical and conceptual investigations.

5.1 Explore

We treat the Explore-step of the VtM model as the empirical and conceptual investigations of VSD, as also proposed by Smits et al. (2019). The Explore-step is supposed to yield an understanding of the values that matter to the actors that matter in investigation. In our article, the exploration of the actors does not come explicitly in this section. We will explore this now.

5.1.1 Finding the Values that Matter

If we followed the VtM approach rigidly, we would probably begin this analysis with defining the potential users and only then derive common values from this group by means of empirical, qualitative investigations. Although this is a good way of discovering some important values, there are some issues with exclusive buttom-up elicitation. According Umbrello and Van de Poel "(...) *any elicited list should be supplemented by principles to ensure that typical AI ethical issues are properly addressed.*"

5.1.2 Values Promoted by the Design

Promoting societal values in the design by orienting the technology towards, not only solving a specific problem and not causing harm, but to make society better can be a difficult concept to translate into design requirements. The Sustainable Development Goals (SDGs) by the United Nations (2012) offers a systematic way to orient a technological development towards something generally 'good'. The third SDG (Good health and well-being) and especially the targets 3.1, 3.2, 3.8, and 3.c relating to maternal and prenatal care as well as training of healthcare workers, have potential to become SDG targets promoted by the design.

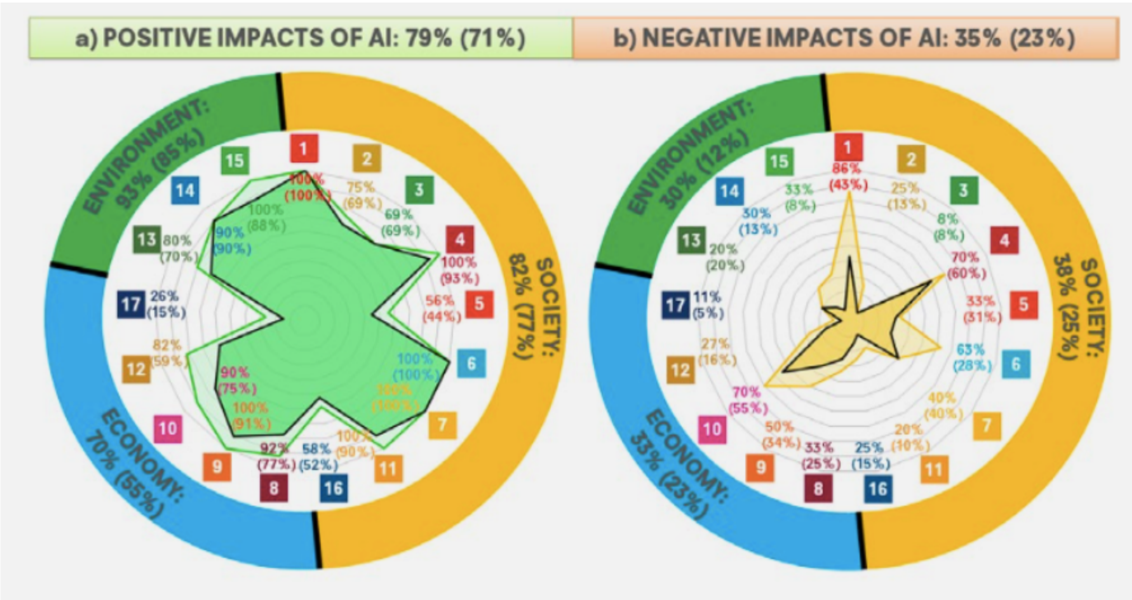


Figure 10: Impacs of AI on the SDGs according to UN (Ke et al., 2021)

UNESCO have analyzed the current uses and potential uses of AI solutions and their impact on the SDG's. The percentages in figure 10 correspond to the impact on the targets of the SDG by AI, positive and negative respectively. UNESCO doesn't see AI as a threat to achieving the third SDG, but neither will AI solve all of SDG3's issues on its current track (Ke et al., 2021). Although they articulate AI's potential to push the ceiling, the biggest promise of AI is tackling unwanted variance in healthcare, substantially raising the average quality of care (Ke et al., 2021). This view on AI in healthcare as a way to reduce errors and heightening the expectation of the overall potential of all care providers to levels traditionally only

achievable by highly specialized, resourceful hospitals fits very well with the aim of the SONAI-project.
In our value framework, we place SDG3 – Good health and well being for all – as the value and a handpicked selection of Targets as design norms. The targets are (United Nations, 2012):

"3.1 – By 2030, reduce the global maternal mortality ratio to less than 70 per 100,000 live births"

"3.2 – By 2030, end preventable deaths of newborns and children under 5 years of age, with all countries aiming to reduce neonatal mortality to at least as low as 12 per 1,000 live births and under-5 mortality to at least as low as 25 per 1,000 live births"

"3.8 – Achieve universal health coverage, including access to quality essential health-care services and access to safe, effective, quality and affordable essential medicines and vaccines for all"

"3.c – Substantially increase health financing and the recruitment, development, training and retention of the health workforce in developing countries, especially in least developed countries and small island developing States"

The SONAI-project aims to improve the quality of care by bringing AI to the medical field of obstetrics such as AI feedback for EFW. This could help to mitigate disparities in obstetric care, especially in hospitals marked by limited access or resources to perform this procedure. In developing countries, rural areas in developed nations or other areas where a specialist can not be expected to be present regularly at the clinic, this AI-solution could be critical in advancing the decision making during pregnancies at risk.

5.1.3 Values Respected by the Design

Drawing once again upon Umbrello and van de Poel, we source a list of values based on the aforementioned ethical principles by the high level EU expert group on AI (HLEG) (European Commission, 2019). These four ethical principles are based on the EU Charter of Fundamental Rights of the European Union (European Commission, 2012) and respecting these principles ensures normative compatibility within the EU ahead of regulative compatibility. The AI HLEG reasons these principles as such:

Human Autonomy: AI systems should be designed to augment and empower humans, not take over decision making power.

Prevention of Harm: AI systems should not increase power gaps in various social configurations, and foreseeable scenarios of misuse must be accommodated before implementation.

Fairness: The costs and benefits of AI systems should be distributed. Likewise bias in training data should not unjustifiably impair freedom of choice. An analysis by an AI should be contestable and not taken for granted.

Explicability: AI systems are as transparent as possible. The capabilities, limitations and purposes are openly communicated.

5.1.4 Finding the Actors that Matter

Determining the actors that matter to this investigation was an interesting exercise. While bringing in values of wider [European] society does contribute to the overall compatibility of AI products, we can't solely design for societal ethics and expect anyone to be able to use the device. Likewise, we can't solely design for the end users' needs and requirements and expect the device to be compatible with the interests of society. Therefore, we need to define the actors that matter – or in VSD language, the direct stakeholders. We define these as obstetricians working in Danish hospitals of all levels of experience, but with a focus on the novices.

5.1.5 Context specific values

By bringing the AI and the user interfaces to the clinicians it was possible to derive context specific values from our identified stakeholders. It is the dynamic exchange of feedback and discussion from our participants that constitute these values. Based on the setup with the participants' review of mockups along with the interviews we were able to articulate how providing good care was a central value that was considered important and what norms that constitute the accomplishment of this value.

Coding our interviews in Nvivo we identified a set of norms that correspond to features needed of medical devices in order to be useful in the clinic. This includes a non-intrusive workflow centered around the patient, a reflection of a clinical certainty regarding accuracy, reliability and effectiveness of medical decisions made within a clinical context. This is also reflected in the professional integrity and accountability between the patient and the clinicians fostering trust and transparency in clinical decisions.

5.2 Conceptualize

At the beginning of the period of this master's project, there were two different functioning UI-concepts which could be used for user-testing and as props for interviews. One was a prototype UI for a working version of the device. This was an engineering prototype version made without much thought to the appearance of screen elements. The other was a UI-mock-up made by Andreas von Essen and members of the SONAI-group during his internship at PCAI.

5.2.1 UI: Prototype

The minimum requirement was that the UI should function as the access point where the clinicians could engage with the AI. This would then be the basis for further research and development. Some of the first goals in the SONAI-project were to develop a functioning AI, build a prototype UI, find a suitable and sustainable system architecture, followed by testing in the clinic. This would mark some of the next phases of development that demanded new insights into how to continue AI development. A new goal was now to look into new UI design requirements based on user centered data. The first engineering prototype can be seen in Fig. 11.

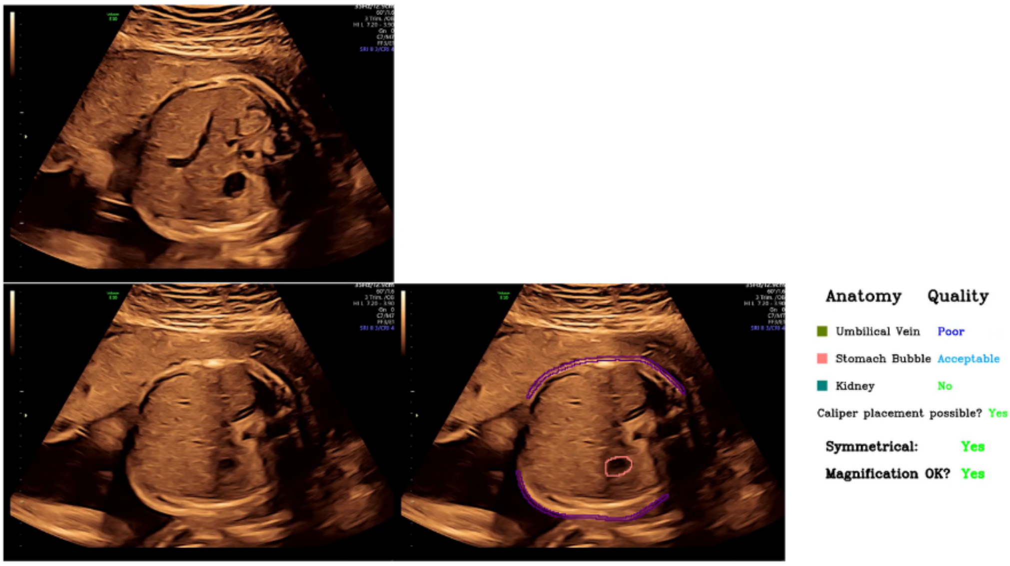


Figure 11: Engineering Prototype of the SONAI AI to EFW

5.2.2 UI: Mock-Up

The engineering prototype AI was not built with final usability in mind, but was again the basis for development of a novel UI. In a discussion with some of the researchers, about further testing and development of the AI in the clinic, we joined forces with two Ph.D students, each with a goal in mind that could be achievable in a collaborative environment. To build a functioning UI-mock-up that was easy to engage with, it was decided to use CANVA as a design platform. This allowed for quick access and edits that we could bring to our participants for feedback. It was decided to build two different iterations, one with minimum feedback and one with maximum feedback to test different functionalities. Since the engineering prototype or the mock-ups was not designed with a value framework at the onset of development, there was still a challenge regarding how to identify these values and translate them into design requirements. To investigate how the AI would mediate the dynamics of the users and the technology we would bring the engineering prototype and the mock-ups to the users for testing. This would allow us to discover requirements based on user feedback that could be analyzed and incorporated into new UI-designs. A selection of some of the UI-mock-up iterations can be seen at Appendix 4, 5, 6, 7.

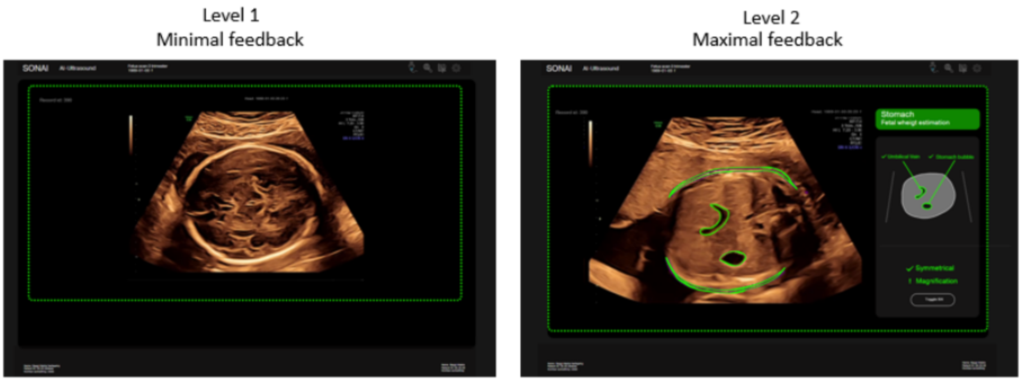


Figure 12: First iteration mock-ups of the AI

5.3 Anticipate

With a conceptualization at hand, we were ready to introduce the prototype and mock-ups to the clinicians. The anticipation section aims to investigate how these two versions transform value experiences, in an effort to reconceptualize the UI elements' mediation of relevant values relating to care and AI.

Value Source	Value
Value Promoted by the Design	SDG 3 Good health and well-being
Values Respected by the Design	Human Autonomy
	Prevention of harm
	Fairness
	Explicability
Context Specific Values	Provide good care
	Receive good care
	Emancipation

The final list of values for the anticipation phase is presented in Table 1 alongside the respective sources. In order to track the translations from values into design requirements, we will add columns to the table showing our reasoning and where in either our theoretical framework or empirical data, a decision stems from. The subsections 5.3.1 through 5.3.4 in this analytical chapter will therefore roughly correspond to each motion of adding a column to this table.

5.3.1 Design Norms

The first column we need to add is the "Design Norm" Column. This column follows van de Poel (2013)'s definition of norms as outlined in section 3.1.1 as prescriptions or restrictions of action. They are the concrete description of what a feature or a decision ought to achieve. Earlier we specified that we placed the SDG 3 targets 3.1, 3.2, 3.8, and 3.c as norms. If the SDGs were written in a VSD vocabulary, the targets might as well have been called "norms" as they are the prescriptions for reaching the specific sustainable development goal. Therefore, in our terminology, the final design ought to contribute to reducing the global maternal mortality ratio (Target 3.1), aim to reduce neonatal mortality (Target 3.2), access to quality essential health-care services (Target 3.8) and boost recruitment, development, training and retention of the health workforce in developing countries (Target 3.c). Therefore, if the result of the design feature ends up fulfilling these design norms, they are indicative of the value being physically present in the design.

The norms for the values respected by the design are, like the values themselves, inspired by Umbrello & van de Poel (2021). Starting with the value of human autonomy, a design norm that can respect this value in AI design is for the AI not to take over too much decision-making power. The value of prevention of harm, can be achieved if an AI solution protects against statistical prejudice. The value of explicability can be achieved by limiting the opaqueness wherever possible, and it should be clear and intelligible why the AI decides what it decides. We interpret the value of fairness as encompassing most of the other values. If we take a step back from an AI-augmented ultrasound procedure, what can ensure that the decision is fair? If the three other values are respected, there should be a mechanism to articulate bias and the decision-making process and only to advise the doctor – not make its own decision. Umbrello and van de Poel translate fairness into justice with reference to the 16th SDG of strong institutions. We are inclined to see the fairness norm as a more of a 'process norm' amplifying and securing the other norms. The processes surrounding the AI to achieve fairness could be a human who is able to be held responsible for the outcomes of an AI-augmented procedure.

Although the above norms relate to AI, we wanted to emphasize that we see the UI of an AI-solution as playing a crucial part in realizing the ethical responsibility of the underlying algorithm. The UI might not have the power to protect against statistical bias in the training data, but there are some ways the UI can articulate the possibility of biased data. We will dive deeper into this in the following sections.

The design norms for the Context Specific Values were derived directly from analysis of our empirical data.

Value Source	Value	Design Norm
Value Promoted by the Design	SDG 3 Good health and well-being	Target 3.1 Reduce maternal mortality
		Target 3.2 Prevent death of newborns and children under 5
		Target 3.8 Achieve universal health coverage
		Target 3.c Secure training of health workforce
Values Respected by the Design	Human Autonomy	AI should not take over too much decision-making power
	Prevention of harm	Protection against statistical prejudice
	Fairness	There should be a human who is able to be held accountable for the AI
	Explicability	AI decisions should be contestable and rationally deductive
Context Specific Values	Provide good care	Indisruptably support the clinical workflow
		Enhance Clinical Certainty
	Receive good care	Strengthen Integrity and Accountability
		Support Patient Engagement
		Increase Trust and transparency
	Emancipation	Facilitate Independent learning

5.3.2 Positive Value Experiences

The next column we need is the Positive Value Experience column. If the concept is shown to hold positive mediation of a value, then we must determine what feature must be kept or reinforced in order for the final product to do the same. Starting with the third SDG, the best way to ensure that the technology has maximum potential for positive change is to ensure that it is as universally usable as possible. Measuring positive value mediation of the third SDG thus requires that we 1) orient ourselves towards culturally neutral design conventions and 2) investigate whether or not our participants find the UI intuitive.

We studied UI design guides for medical devices, AI, and universability (e.g. Lieberman 2009, Murphy, 2018) and cross referenced with both the prototype and the mock-up. We observed three medical students use the AI on a pregnant volunteer, one without feedback, one with minimum and one with maximum. In general, even the participants who had little-to-no prior introduction to the procedure had a degree of success with both the minimum and maximum feedback version of the AI. We see this as an indication that even in its current conception, the AI does show potential for independent learning and the inclusion of unspecialized clinical staff in a highly specialized environment. The UI elements allow a novice to get a sense of their performance, whether they are finding standard planes or not, even without a senior clinician providing them with feedback: "*Finally we can contribute as young doctors, the old doctors doesn't have the time to wait for my unsteady hands*" (P1). By utilizing the potential learning value the AI could support training of a healthcare workforce that supports the third SDG promoting well-being for all ages and especially the targets surrounding maternal and newborn health.

We distinguish between the context-specific values of providing good care and receiving good care in order to emphasize where the primary impacts of the design requirements lie. In practice these are intertwined, and we mostly sourced both sets of values from clinicians, although we did speak to a pregnant volunteer who agreed to act as a patient for our participants. The "patient" expressed an interest and agreement that an ultrasound procedure could be a place for AI technologies: "*AI could be a tool that supports their decision making. It is a good way to become better at the clinic faster!*" (P5).

When AI is introduced as a novel technology in the clinic when performing EFW it should be on a level that is not esoteric, increasing the distance between the patient and the clinician. The UI elements presented to the patient should therefore be explicable to the patients or their relatives as non-clinicians and should not be up for misinterpretation that could lead to distrust. We use this encounter as an example of a positive value experience as the patient was additionally able to understand that the UI elements served as guidelines for the clinician and not as interpretations on the baby's health. Further development of the AI-UI with increased patient involvement should focus on confirming this, but the patient experience was not directly in the scope of this project.

<table>
<tr><th>…</th><th>Value</th><th>Design Norm</th><th>Positive Value Experience</th></tr>
<tr><td rowspan="4">…</td><td rowspan="4">SDG 3 Good health and well-being</td><td>Target 3.1
Reduce maternal mortality</td><td rowspan="2">The technical and professional abilities in the healthcare setting enables identifying and responding appropriately to a EFW</td></tr>
<tr><td>Target 3.2
Prevent death of newborns and children under 5</td></tr>
<tr><td>Target 3.8
Achieve universal health coverage</td><td rowspan="2">The device make sense in theory, training and practice</td></tr>
<tr><td>Target 3.c
Secure training of health workforce</td></tr>
<tr><td>…</td><td>Human Autonomy</td><td>AI should not take over too much decision- making power</td><td rowspan="2">The device augments the practitioner’s decision making when taking optimal pictures for accurate EFW.</td></tr>
<tr><td>…</td><td>Prevention of harm</td><td>Protection against statistical prejudice</td></tr>
<tr><td>…</td><td>Fairness</td><td>There should be a human who is able to be held accountable for the AI</td><td rowspan="2">- Accountability is traceable and on site.
- The reasoning behind AI-interpretations are transparent.</td></tr>
<tr><td>…</td><td>Explicability</td><td>AI decisions should be contestable and rationally deductive</td></tr>
<tr><td rowspan="2">…</td><td rowspan="2">Provide good care</td><td>Indisruptably support the clinical workflow</td><td>- The AI-interpretation is considered as accurate
- Minimal interruptions of clinical workflow.</td></tr>
<tr><td>Enhance Clinical Certainty</td><td>The device improves confidence in decision-making.</td></tr>
<tr><td rowspan="3">…</td><td rowspan="3">Receive good care</td><td>Strengthen Integrity and Accountability</td><td>The patient can trace and understand the responsibilities</td></tr>
<tr><td>Support Patient Engagement</td><td>The patient experiences that their input is used and respected.</td></tr>
<tr><td>Increase Trust and transparency</td><td>- The patient can trust the clinical decisions based on interpretations that include AI in diagnostics.</td></tr>
<tr><td>…</td><td>Emancipation</td><td>Facilitate Independent learning</td><td>The development of skills can be controlled and managed by the clinician on their own.</td></tr>
</table>

5.3.3 Negative Value Experience

The term 'negative value experiences' refers to the technology's negative mediation of a value. During the user-testing scans, some of the clinicians would solely focus on the feedback from AI, trying to optimize the scans for full green feedback and not so much if the ultrasound picture were accurate enough to be used to EFW. Although the AI is designed to be indicative of the image quality, our participants appeared to not look for standard planes, but rather to make the UI turn green. This poses an interesting dilemma between the value sources. For maximizing universability of the UI in all contexts to enhance its ability to do the maximum good for maternal and newborn health, would it then not be a good thing that the UI is so intuitive that the underlying medical specialist knowledge phases into the background? Although this is desirable for places with no alternative and a lack of medical specialist knowledge, it might not be desirable in all contexts to remove the need for obstetricians to learn how to perform EFW with no AI assistance. One observation was how a participant would uncritically freeze images based on feedback from the AI:

"*I think it's a bit hard, since I don't really have any background information on this, like, I have to be confident myself, and then I need approval from the AI tool, with the AI I would feel more confident, but right now I'm not confident at all, and then it says, you're probably right, and then I'm like, okay this is as good as it gets...*" (P6)

It is important that the AI augments the practitioner's decision-making by providing more context and strengthening certainty, and although a clinically approved tool can reasonably be expected to enhance the procedure, it does provide accountability issues if its analysis in hindsight led to the wrong intervention. In its current minimum-feedback conceptualization the AI-UI is indisputably 'green' when in a standard plane and 'red' when not placed correctly. This way the AI-feedback can be said to dictate the workflow of the operator and is thus negatively mediating the values of explicability, human autonomy and providing good care (specifically the design norm of clinical certainty). Another place where the technology had a contestable mediating impact was through cognitive load affecting the clinical workflow.

Additional screen elements and variables that need to be checked could potentially over-complicate the procedure, counter-intuitively taking up even more of the clinician's mental capacity than without the AI support system. Initial reaction-time testing performed by our supervisor from the SONAI-group on three medical students showed that the latency between hearing a sound and pressing a button while simultaneously performing the scan increased considerably the more screen elements were present. However, talking to the participants they also expressed initial enhanced confidence and elevated clinical certainty when performing EFW with the maximum-feedback UI. The feedback provided was perceived as trustworthy for clinical decision making and provided reassurance to the operators that they were in the correct plane, thereby reinforcing the operators confidence whether they discovered the correct plane or not.

5.3.4 Design Requirements

The last column we need to add is the Design Requirements-column. Smits et al. write in their use of the VtM model that "*where the technology held negative mediation, a design requirement was created.*" (Smits et al., 2022, p. 48). When defining design requirements based on both theoretical and empirically sourced values, we noticed an interesting value clash between the values respected by the design and the context specific values. If we relate the context specific values, to the four values sourced from the AI HLEG, a minimum feedback version is in a direct value clash with all four. In line with these, an AI should not take over the decision-making of the doctor, or the inclusion of the patient. The ultrasound images required to perform EFW need to be explicable, confirming exactly why a certain standard plane is adequate for clinical decision making. By not displaying in the UI why the AI claims that the picture is good enough for an accurate weight estimation, the AI appears both deceptive, authoritative and insensitive towards its own bias if used in a situation it is not trained for.

…	Value	Design Norm	…	Negative Value Experience	Design Requirement
…	SDG 3 Good health and well-being	Target 3.1 Reduce maternal mortality	…	Preventable complications are not accommodated. Unnecessary intrusive interventions are concluded based on the AI-feedback.	UI that helps with identification of standard planes when performing EFW. The AI can help with clearly identifying anatomy that indicates a wrong standard plane as well as a correct standard plane.
		Target 3.2 Prevent death of newborns and children under 5			
		Target 3.8 Achieve universal health coverage	…	The practices surrounding / required by the device doesn't match the context of the clinic.	Accessible and understandable UI with instructions that are intuitive to an unspecialized healthcare professional
		Target 3.c Secure training of health workforce			
…	Human Autonomy	AI should not take over too much decision- making power	…	The practitioner 'games' the scan, looking for all parameters to 'turn green'.	- The AI does not output an actionable analysis based on the scan. - Help to flag ambiguous cases where expert assistance is needed. - An indication of how sure the AI is of having found the correct SP e.g. a per- centage indicator to avoid prejudice bias.
…	Prevention of harm	Protection against statistical prejudice			
…	Fairness	There should be a human who is able to be held accountable for the AI	…	- Accountability and responsibility is symbolic - the AI's analyses are 'taken for granted as true'. - The AI-reasoning is considered opaque.	- The AI's maximum information layout should provide a rationally sound deductible reasoning for its estimations, that can be intellectually contested by a human operator. - By default the 'maximum information' layout is turned on.
…	Explic-ability	AI decisions should be contestable and rationally deductive			
…	Provide good quality care	Indisruptably support the clinical workflow	…	Inaccurate handling of the US machine when the AI can't keep up with real-time	- Provide a visual indication of the current latency of the AI
		Enhance Clinical Certainty	…	The device makes false positives	An indication of how sure the AI is of having found the correct SP e.g. a percentage indicator.

…:	Receive good care	Strengthen Integrity and Accountability	…	Fear that the AI makes an analysis that is 'mediated' by the Doctor. No human eyes on the test results.	The UI should show the logic of an analysis, and the logic should clearly correspond to medical literature.
		Support Patient Engagement	…	Worry that the AI interprets on the health of the baby.	Misinterpretable UI elements should be limited or changed.
		Increase Trust and transparency	…	The device is too esoteric and increases the distance between patient and practitioner.	- The UI should also be at the same 'height' as the pregnant person. - Interpretations of UI-elements from the patient should not lead to distrust in decisions from the clinicians.
…	Emancipat ion	Facilitate Independent learning	…	The practitioner gets very good at performing scans with the AI, but becomes over-reliant on it.	- Toggle-able screen elements for practicing without "safety wheels" - Directional support system when performing EFW

5.3.5 Example: Negative Mediation of Clinical Certainty

One of the participants expressed that if the device is present in their department, then "*I have to trust that it works (...) and it must be approved and validated.*" (P5). The doctor in this case expressed a trust towards the approving bodies and regulation in place to ensure that medical devices only get implemented if they are tested and approved for clinical use. We interpret this as an expression of a need to not treat this technology differently because it is AI. If a non-AI medical device would not show specific UI elements, why should the AI do so? Then it must not be reliable, and if he can't trust that it works, it becomes noise:

"*Machine learning and such... I am totally overloaded already with my own stuff. I just want [a tool] that is easy, good and functions – like all the other tools.*" (P5)

Although training and education is associated with the use of a new device in the clinic, in an information heavy environment, it would be problematic if every user had to assess whether or not a given tool is AI-trained or not, when using it for medical decisions. This participant expressed an expectation that you can trust the analysis of the devices without understanding its technical inner workings:

"*I am 100% going to trust the AI-interpretations.*" (P5)

This notion is interesting when held up against the minimum feedback version of the UI relating to the values of human autonomy and explicability. If the information and reasoning behind the AI's analysis is 'hidden' behind e.g. a color indication, some of the responsibility leaves the department. It becomes opaque and shared between the clinician performing the procedure, the doctor in charge of the department, the approval board that approved the device and even the designers of the technology. Although it is a difficult decision, we ultimately decided that the negative mediation of the HLEG values – especially that of explicability – weighed heavier than the negative mediation of the individual doctor's workflow and clinical certainty by displaying more information than might be necessary had the underlying technology not been AI.

5.4 Re-conceptualize

With a filled-out table containing a list of design requirements in hand, we're able to turn them into a new concept – this time in the form of a re-conceptualization of our UI mock-ups. These will act as the blueprint that should be used for a new iteration of a working UI-prototype. The design sessions were conducted in CANVA. We designed to accommodate that a design requirement should be created if the concept holds negative mediation when introduced to the participants.

5.4.1 Universability

Example: Figure 13: Femur Standard Plane – Maximum Feedback – Correct Placement

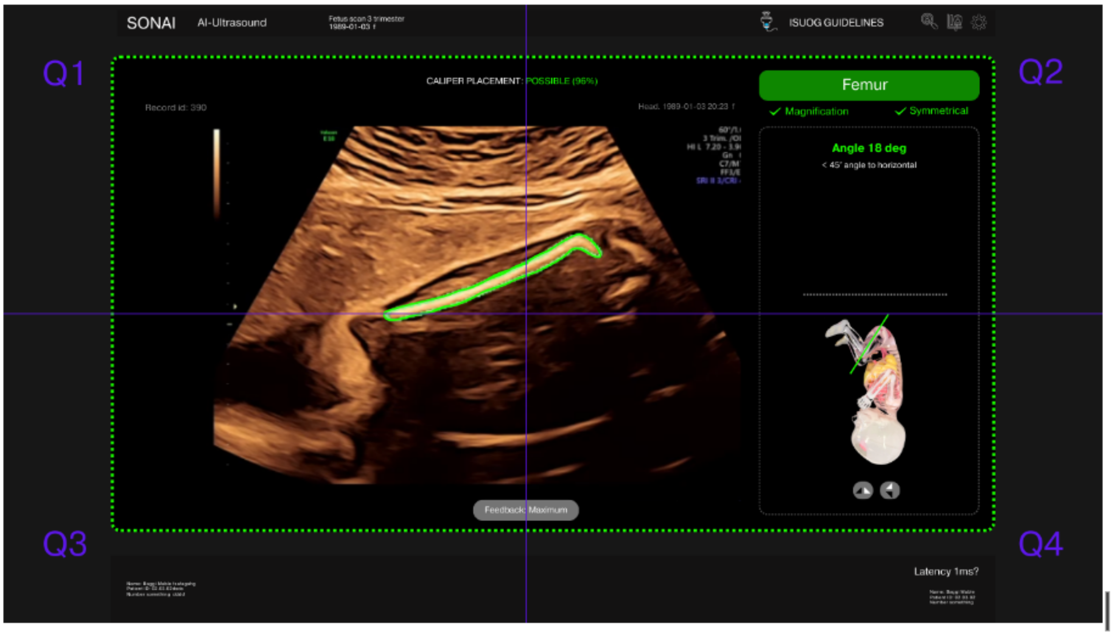


Figure 13: New Design: Maximum feedback of a Femur Standard Plane

We looked into to the UI design to make a more accessible and culturally neutral UI that will not only make the AI more intuitive to use, but increase its overall capacity to “Ensure healthy lives and promote well-being (...)” (United Nations, 2012). A specific point of interest was the UI-elements related to technical terms and abbreviations, these had to be aligned from a clinical standpoint. The selected technical terms are directly from the ISUOG (2019) Practice Guidelines for ultrasound assessment of fetal biometry and growth, which is regarded as standard within the medical field of obstetrics. One of the Ph.D’s in medicine was supervising the mock-up development and participated in the process of selecting these technical terms and abbreviations. She had many years of experience in the field of obstetric and specifically in performing EFW. The insights she provided was valuable as she could confirm or disregard UI-elements and formulations from a clinicians point of view. Abbreviations and acronyms are a daily part of clinical jargon and contribute to more efficient flow of critical information in highly specialized medical departments. The UI-elements have to align with the practice in question and not confuse the clinician or complicate procedures with new abbreviations, which could pose a threat to patient safety Tariq & Sharma (2024). This also holds true for colors or symbols or shapes that could be misinterpreted. The use of the AI has to be intuitive and understandable to the operator to function at capacity, the utility of the AI could otherwise be disregarded. The objective of the AI is that it would help the operator to confirm the criteria needed for evaluation of the biometric images, by illustrating each fulfilled criterion. Here we decided upon the color green as a color identification. The reasoning is that If the operator reaches a standard plane, everything is green, good and understandable. For example the ultrasound display in Fig 13 is encompassed by a dotted green frame, the word "Femur" appears in the box that indicates which plane is at hand in Q2. The checklist feature in Q2, also underlined by the ISUOG guidelines (2019) and delivers a rationally deductive explanation by providing in situ illustrations of how it achieved its analysis of the ultrasound images. Overall the UI elements were selected to be understandable to the users in question and is based upon feedback from those who have to use the AI.

5.4.2 Accountability

Example: Figure 14: Head Standard Plane – Minimum Feedback – Correct Placement

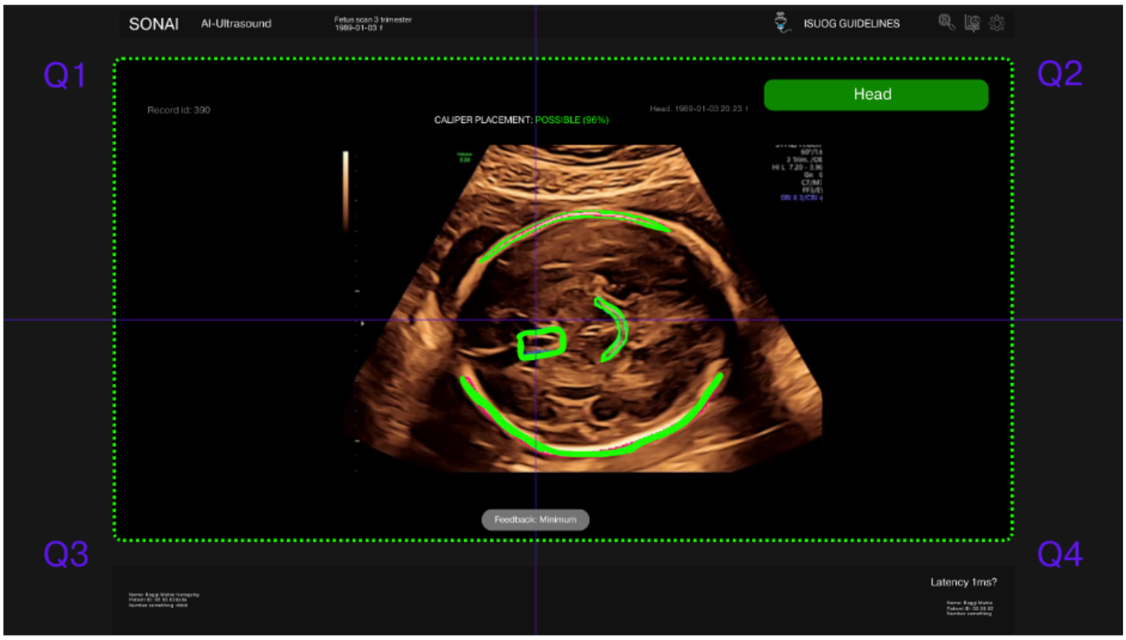


Figure 14: New Design: Minimum feedback of a head Standard Plane

A returning discussion about the UI-elements was that of how much feedback is needed to be portrayed in the UI-elements. Too much feedback could be considered abundant, while too little feedback unusable. This was the initial idea behind the two iterations of the mock-ups, one with minimum feedback and one with maximum feedback. We did however discover that a minimum feedback version could be contradictory in the sense that it could lead to an accountability problem. To mitigate this we propose the maximum feedback version as a default and created a toggleable UI-element for the level of feedback in the bottom of the frame at the intersection of Q3 and Q4. This would allow the clinician to make a choice about the feedback they deem necessary.

5.4.3 Usability

Example: Figure 15: No Standard Plane – Maximum Feedback

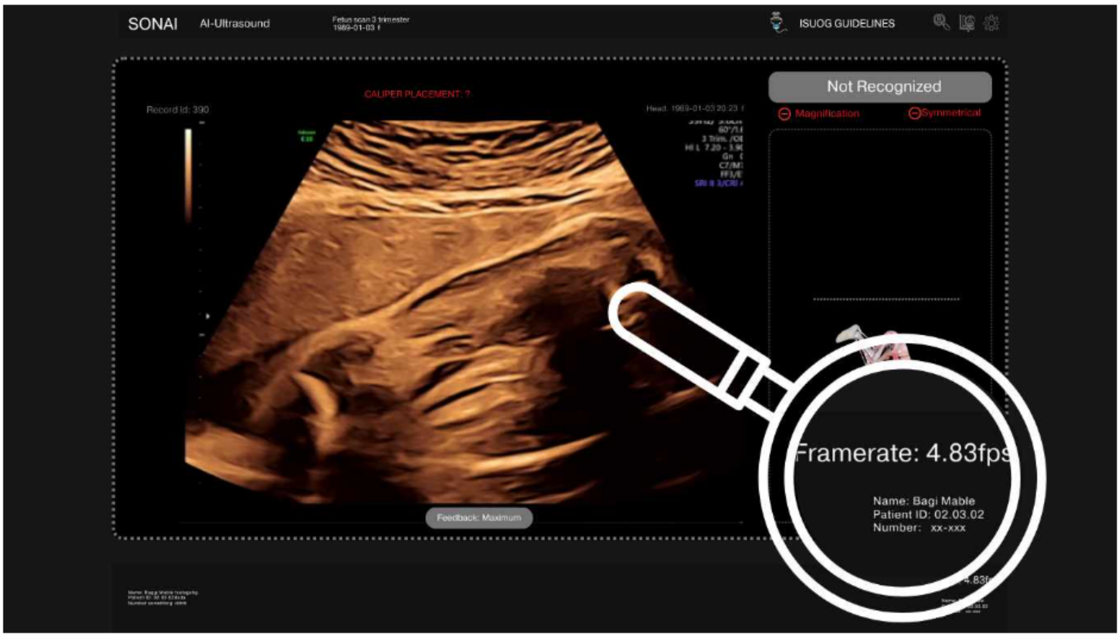


Figure 15: New Design: Maximum feedback No Standard Plane

During testing of the AI a delay in the feedback was observed. This is a negative value experience, compromising clinical certainty, disrupting workflow, and limiting the ability to deliver timely decisions during the procedure. The optimization of the AI performance is preferable, yet we also suggest an improvement to the UI. An indication of how fast the image is being generated would be preferable. The framerate of the AI-image dropped at different points, sometimes related to adjusting settings on the ultrasound machine. Therefore, a framerate counter in the bottom corner of Q4 could be considered, with a color coding indicating optimal conditions, in order to give the clinician more control over the machine's settings and finding an optimal balance that works for them.

5.4.4 Orientability

Example: Figure 16: Stomach Standard Plane – Maximum Feedback – Wrong Placement

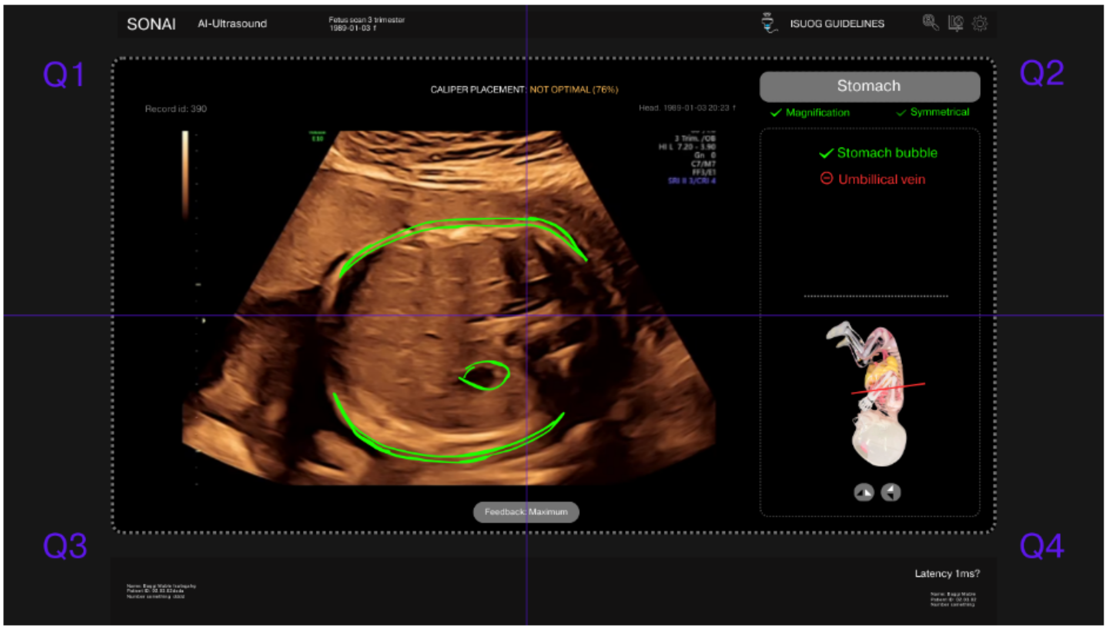


Figure 16: New Design: Maximum Feedback Stomach Standard Plane

The feedback from a checklist could help to assure identification of a standard plane with the corresponding anatomical landmarks. This feature was an iteration taken from an engineering prototype and copied into canva. The Labels stay the same and turn green when a correct anatomical landmark is identified. Like in this example where the stomach bubble is in the picture. Moreover, the participants asked for a sort of guidance system that can indicate how to locate the next SP needed for EFW, for example by recommending probe manipulation or changing settings on the ultrasound machine. In Q4 we suggest an avatar to help confirm if the operator has located a correct plane by indicating with a colored confirmation line. In this case a red line illustrating that the scan is not in the correct plane.

6 Discussion

6.1 Ethical Design Reflections

AI is not at all a new subject of interest to VSD researchers. A number of the early works using VSD were concerned with AI and the morals of computers. Friedman & Kahn (1992) investigate the possibility of computers to become moral agents. In order to be able to be held morally responsible for the consequences of an action, there must be some sort of intentionality behind the act. Intentionality was introduced into philosophy by Franz Brentano in the latter part of the 19th century (Jacob, 2023) and according to Searle (1980) it describes the relationship between the mind wanting something and the ability to satisfy this by action. Intentionality in humans is important, especially in relation to two of its counterparts, the accident and the instinct. If there was no difference between the consequences of an accidental act and an intentional act, then moral and legal systems would struggle to appropriately assign responsibility and culpability.

6.1.1 The Chinese Room

Searle argues against a thinking AI based on the notion that a computer merely manipulates symbols but has no way of attaching meaning to what it is doing. He argues against the possibility of 'strong AI', that is AI that the right software program will be able to provide a computer with a mind (Searle, 1980). As a digital computer according to Searle is only able to take an input and process it to an output based on defined rules, it is incapable of understanding what its actions 'mean':
"Consider a language you don't understand. In my case, I do not understand Chinese. To me Chinese writing looks like so many meaningless squiggles. Now suppose I am placed in a room containing baskets full of Chinese symbols. Suppose also that I am given a rule book in English for matching Chinese symbols with other Chinese symbols. (...) Imagine that people outside the room who understand Chinese hand in small bunches of symbols and that in response I manipulate the symbols according to the rule book and hand back more small bunches of symbols. Now, the rule book is the "computer program." The people who wrote it are "programmers," and I am the "computer." The baskets full of symbols are the "data base," the small

bunches that are handed in to me are “questions” and the bunches I then hand out are “answers.” Now suppose that the rule book is written in such a way that my “answers” to the “questions” are indistinguishable from those of a native Chinese speaker. (...) I satisfy the Turing test for understanding Chinese. All the same, I am totally ignorant of Chinese. And there is no way I could come to understand Chinese in the system as described, since there is no way that I can learn the meanings of any of the symbols. Like a computer, I manipulate symbols, but I attach no meaning to the symbols.” (Searle, 1980)

The “Chinese room argument” is a commonly referenced – and criticized – thought experiment aimed to disprove the notion of computers being able to develop a mind. Although a lot has happened in the fields of computing and especially AI since 1990, it can still be used to reject intentionality in AI systems of today, such as ML. For example, natural language processing (NLP) models such as GPT-4 which is the current basis of OpenAI’s ChatGPT, embed words as vectors with up to thousands of dimensions. The position of each word’s vector in this high-dimensional space encodes its meaning relative to other words by means of proximity. In an over-simplified sense, this means that averaging the vectors of “Germany”, “Olaf Schultz”, “France”, “Emmanuel Macron” and “Denmark” should provide the approximate coordinates for the vector for “Mette Frederiksen”.

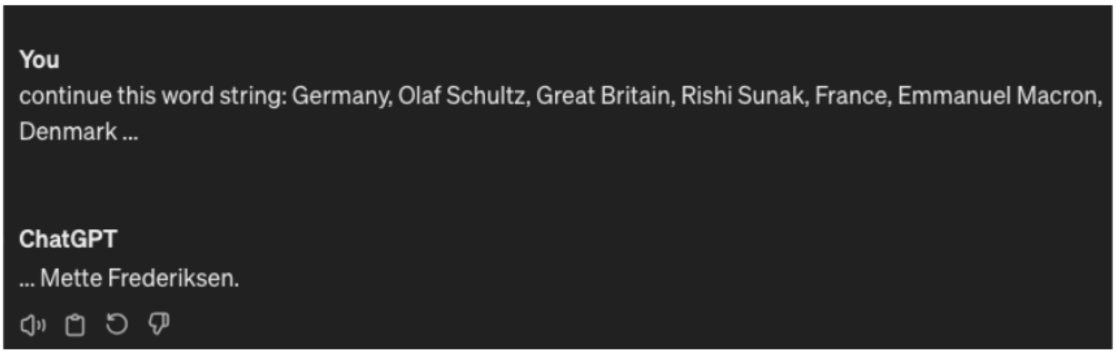


Figure 17: ChatGPT Promt

There are obviously other factors at play in generating a ChatGPT response beyond vector arithmetics, but the point is to illustrate how the model generates a response based on statistical patterns rather than deep semantic understandings, much like the Chinese room argument. Relating technological meditation and post-phenomenology to the concept of intentionality, according to Ihde (1990) the way humans act within the world is not solely a property of human consciousness. It is distributed across the human-technology-world relation. Ihde describes the opacity of the technology as a point of interest when determining a human's ability to act upon the world through the technology. He calls an opaque connection an "enigma". In the case of hermeneutic relations, this sets some requirements for the reliability of the technology that inscribes a worldly phenomenon onto graphs. If a technology interprets upon something that we otherwise do not have access to see or the ability to comprehend, how much of an opacity is there and what other factors can cause an instrument to show a given reading? The mediating role of technology means that even if it cannot hold moral agency in itself, our ability to be moral agents in the world is tied to it and its opacity – its enigma. So what does this mean in a VSD sense? Even if AI – or any sophisticated computer system – can submit human-like answers to questions faster and with a greater accuracy than a human, it cannot be held responsible for the consequences of the action. Much like a heavily bureaucratized institution, the responsibility is diffused, and even if a symbolic head will take the fall after a scandalous case, no one is directly to blame. It's the fault of the system. Friedman & Kahn (1992) conclude that a human should not be reduced to a machine-like status, fulfilling a mechanical task to serve the computer system. Moreover, a computer should never impersonate a human being. These design approaches both seek to place the decision-making and responsibility with a human operator.

6.1.2 Responsible Design

The experience of us humans as subjects in the world is radically transformed when mediated by technology (Kudina & Verbeek, 2019). Technologies like ultrasound play a role in moral decisions especially in regard to pregnancy. Ultrasound machines transform the concept of pregnancy from a condition you can be in to mediating a potential patient, thus changing the way expecting parents relate to the fetus (e.g. as an unborn child). This experience can influence moral decisions such as whether to terminate the pregnancy or making specific health related interventions. In this way the ultrasound machine shapes ethical perspectives and decisions through technological mediation. Now that AI is introduced into the interpretation of ultrasound images it is involved in the decision making process thereby mediating this moral subjectivity.. Even though the clinical decision making and responsibility falls upon the clinician, if they trust that the AI-feedback is the best course of action, they might decide upon the AI-interpretation as sufficient. This highlights the need for ethical reflection and consideration when developing technologies like AI. With VSD it was possible to take the ethical consideration into account and translate these values into design requirements. Technologies can not just be regarded as neutral tools but have individual, social, and societal implications: technologies shape the behavior, experiences, and even moral frameworks of the people who use them. This also underlines designer's responsibility in shaping a technology that aligns with values, by materializing them into the design process. In this study we have tried to incorporate these values into the design of the UI and we have been testing the AI alongside the potential users. This has allowed for an empirically based anticipation of how the technology might have a mediating impact on the value of the actors before the technology has been fully included in the clinic.

6.2 AI-UI in a Design Process

In this study we have identified and defined requirements for a UI that aligns with the needs in the clinic and societal values. Our re-conceptualization can act as a blueprint for future integration into a working UI, this still needs to be designed. To embody our requirements into a working design and to examine if this design actually lives up to the demands of the clinic. A methodological approach to UI-AI development and testing could be beneficial. A modified version of the V-model could be used as a structured framework for the development stages. The model consists of two legs forming a V. The left leg is known as the project definition phase, while the right leg constitutes the testing phase. In the project testing phase, the testing helps to confirm the design and define the requirements needed in development. The verification steps confirm whether the design corresponds with the initial requirements. If a step in the verification fails the development process, the design process can move backward by redefining the initial requirements needed for the design definitions.

With this model it is possible to insert our requirement analysis as a project definition and linearly develop and validate the requirements to ensure that the design requirements are systematically addressed. This can help to ensure a methodological development with the integration of our design requirements into a working UI design, while reducing the risk of overlooking critical aspects of the UI design. When we propose a modified version of the V-model as a methodological approach for UI-design development we are concerned with ensuring a flexibility in the design process. We regard the V-model as a rigid process of development that may have some difficulties aligning with the nature of the inner workings of the SONAI-project, where rigidity and structured development could pose a potential challenge. By combining the V-model's rigor of validating the requirements in the design process by involving those who have to use the AI, with more iterative methodologies for validation and development. A more balanced solution could be lucrative. This would ensure both through validation of the design requirements as well as the flexibility needed for a more effective UI-development.

6.3 Clinical AI as a Tool for Learning

A central discussion and yet-to-be-determined factor about the SONAI-system is the intended area of application. Put simply, it is not clear at the moment whether or not the AI is supposed to live in the clinic as a tool for the unspecialized – or even specialized obstetrician or as a learning tool for trainees. Stuart Dreyfus and Hubert Dreyfus suggest that someone who wishes to acquire a new skill can do this via one of two paths – either they can start practicing and adjust based on trial and error, or they can seek to build upon the knowledge and experience of others by picking up a manual or seeking an instructor (Dreyfus & Dreyfus, 1980). The problem related to the AI of our investigation however is that due to the clinical nature of the area of application, neither approach is necessarily effective. Trial and error would require an instructor or senior doctor to qualify the decision, as avoidable diagnostic error is unacceptable as a means for learning. Conversely, it is impractical to have an instructor qualify every decision or do one-on-one training, as that poses unrealistic resource requirements. Dreyfus and Dreyfus have categorized the path from novice to expert into five stages: Novice, Advanced Beginner, Competent, Proficient, Expert. We will not break down every stage, but in its essence, the model describes how a learner might handle a problem within the skill they are acquiring. A novice language learner would think about the grammar of that language, trying to remember the correct order of the words when practicing in front of a teacher, mirror image or language learning app avatar. One who is an expert in a second language has transcended the rules and has an intuitive grasp of a situation. They are not checking a dictionary or practicing sentences but would be able to speak their mind and react to a new situation they have never encountered before in that language. Building up this intuition is at the core of the problem to be solved by the SONAI-system. The question is whether or not using the AI will help a learning clinician become an expert on EFW scans? Our proposed UI for the AI with the maximum amount of elements could be a valuable tool for the novice. A novice relying on rules would need to understand why they are doing something wrong or correct. A trial-and-error approach can be accommodated via the color coding and percentage indicators while the UI being based on the ISUOG guidelines will accommodate a more scholarly approach. Since the UI clearly states which elements are visible and at times provides a number indication of how clear they are, this version should help the learner approach competency. An active learner who wants to get better at performing the EFW might then turn to the minimum feedback version. As the explicit reliance on rules and codes becomes less of a help and more of a cognitive noise, the reduction of screen elements also forces the clinician to obtain a sense of how to maneuver the probe. The obstetrician on our project has described her mastery of handling the probe to locate standard planes as if she has a mental image of the fetus, its location and how to navigate to every standard plane after a few seconds of 'calibration'. The opposite argument can also be made – this AI will not help the general clinician obtain the skill, because the mastery of it has become redundant. The UI turns green when it's in a standard plane, and red when it's not. The operator would require only a checklist and foundational knowledge to perform the procedure, and will grow their proficiency at making the red bubble turn green rather than locating a standard plane. The sonographers would still be around – someone needs to be able to help with standard deviating cases. But for the average obstetrician the AI successfully solves the problem without giving the clinician the ability to form a mental image of the fetus during the procedure. It is not up to us to decide or try to give an answer on whether this is good for the field or the patient or not. Technology makes skills obsolete while it creates new skills.

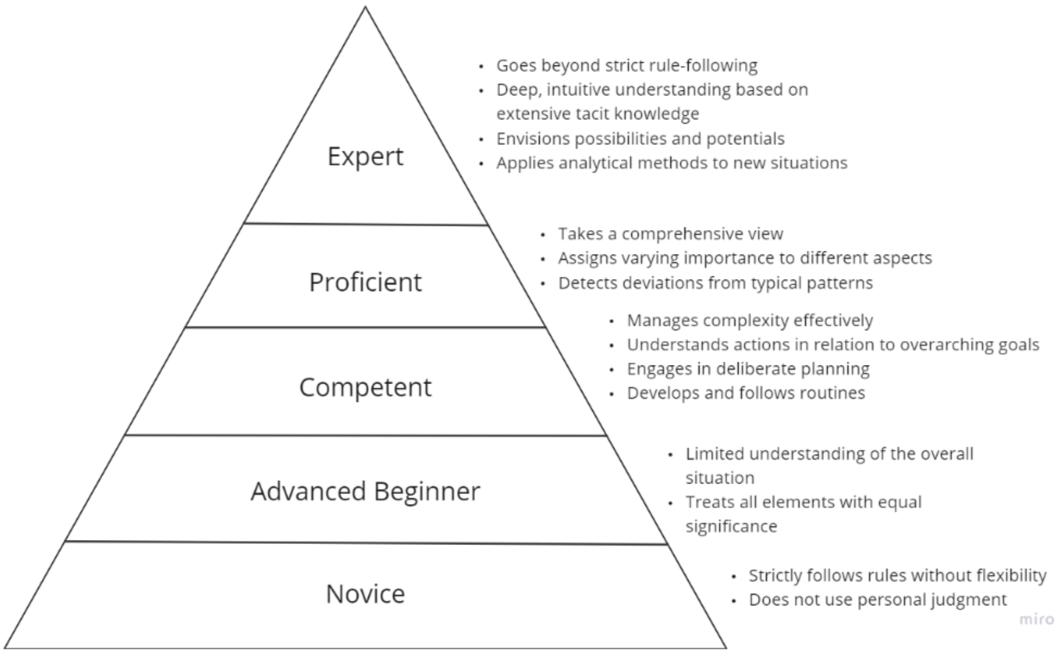


Figure 18: Visual interpretation of Dreyfus & Dreyfus' Model (Dreyfus & Dreyfus, 1980) inspired by secondary literature

6.4 Techno-Anthropological reflections

6.4.1 Researcher, Student or Consultant?

A usual problem that can arise when doing external collaborations as students working on a project is that of balancing the study of the technology in question with the degree of involvement into the decisions regarding its further development. Torben Elgaard Jensen explores this in his 2012 article *Intervention by Invitation: New Concerns and New Versions of the User in STS.* Investigating science and technology development from the outside has long been a privilege of Science, Technology and Society (STS) researchers, a category of theories and methodologies that techno-anthropology more often than not finds itself relying upon. However in recent times, STS-researchers are being increasingly included in larger projects, being challenged to use their theories and methods about the present user-situation to intervene in decision-making and propose paths for better future outcomes. Entering into the collaboration with the SONAI-group posed an interesting challenge, because in addition to helping them develop their technology, they also expected us to generate research. The Pioneer Center for AI (PCAI) is funded via the participating universities and our partners are thus researchers or Ph.D.'s themselves. This meant that not only did we have to find the balance between external consultants and students at a university, but also balance two parallel studies of the AI - a study as associated master's students of PCAI as an institution and a study as master's students enrolled on AAU as Techno-Anthropologists.

6.4.2 On Interdisciplinarity

Our study connected previously unconnected actors - the users, the AI HLEG, and the designers - but also indirectly opened up for more discussions after our departure. As Jensen puts it, we were invited to invite other actors into the study (Jensen, 2012). The normative framework from VSD further articulated indirect stakeholders and although we did not do a patient-study, we started the discussion by designing for the value of 'Receiving good care'. A user interface is the materialized connection between the world of the designer and the world of the user, so giving AI-UI this attention and formulating design requirements that are traceable into empirical findings derived from social science has the potential to leave a lasting impact. Tom Børsen proposed a model of the techno-anthropological field as a way to orient and examine a techno-anthropological study (See Fig 19). The points of the triangle represent the different areas of investigation, and the sides represent the techno-anthropological expertise and methods to connect these areas.

Interactional expertise as a term originates from Collins and Evans 2003 paper "*Constructing Expertise: In a Third Wave of Science Studies?*" who describes it as the mastery of the language of a domain, without the mastery of said domain. Our investigation bridges the gap between the clinicians (users) and the SONAI team (experts) via the interactional expertise we acquire by diving deep into the life world of the user using qualitative methods (Børsen, 2013).

Social responsibility as a competence refers to the Techno-anthropologists ability

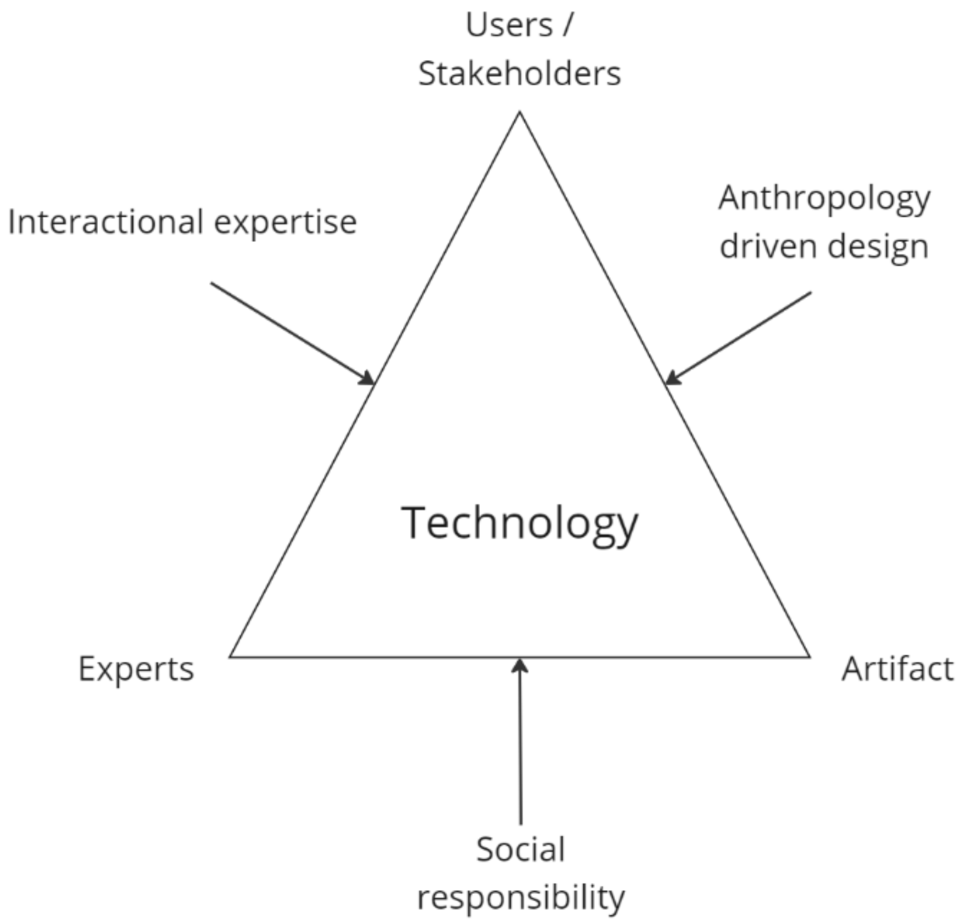


Figure 19: The Techno-anthropological Triangle (Børsen, 2013)

to be able to navigate ethical dilemmas. An ethical dilemma arises when one or more norms or values are in conflict. For example, the four EU ethical principles are not expected to have an equal saying in every issue, and should be weighed against each other when there are conflicts. In healthcare technology, the value of prevention of harm is likely to outweigh that of human autonomy e.g. via constant video surveillance of patients at high risk or the use of patient data to train ML models for better diagnosis. In law enforcement technologies however, human autonomy is likely to outweigh prevention of harm in the same technology examples. Constant video surveillance of potential criminals causes massive privacy concerns and training ML models on historical criminal data can lock in bias and unfairly target specific demographic groups. All this to say that there isn't one answer to correctly weighing these issues, but someone needs to make this assessment and as an interdisciplinary ressource on a project who investigates the user group, chance is the Techno-Anthropologist will have an impact on this.

Anthropology-driven design refers to the combination of participatory design approaches with anthropological methods. The idea is, that what is important in a technology, is how it is used in the intended environment. You do not get an understanding of how the AI will be used by looking at the specifications of it – you can only get that by understanding current practices and include the potential users in the design process. Our creole VSD/VtM methodology encompasses all sides of the triangle. VtM provides a framework for anthropology-driven design with its iterative approach to prototyping and testing with user groups. The AI specific VSD-methodology of Umbrello and van de Poel contributes with a vocabulary for designing in a socially responsible way. Our interactional expertise comes into play directly in the project as we are explicitly investigating the interplay between the realm of the designers (experts) and the realm of the potential users.

7 Conclusion

Artificial intelligence (AI) in the form of machine learning is here to stay and is rapidly entering many fields. For medical technologies, the possibilities seem endless and the machines can get very good at performing analyses usually reserved to human clinicians, but our study showed that there is a conversation to be had about the interface between the operator and the inner workings. Designing user interfaces for AI technologies in health care as masters students on a social scientific programme is an interdisciplinary exercise requiring many technical conversations and a solid theoretical framework. We used value sensitive design and adapted our approach based on an AI-specific framework to value identification and a post-phenomenological methodology for translating values into design requirements. The values of importance were the third Sustainable Development Goal, human autonomy, prevention of harm, fairness, explicability, providing good care, receiving good care, and emancipation. We translated these values into design requirements promoting context-free and universal usage, decreasing opaqueness of the decision-making, improving accountability, and reinforcing novice obstetricians' ability to learn independently. The process involved designing UI elements that align with clinical standards, selecting technical terms and abbreviations from authoritative guidelines, and incorporating feedback from experienced clinicians. Key features such as color-coded feedback, a toggleable feedback system, and an avatar for orientability were included to improve the operator's experience. The design changes we propose are relatively minor in the sense that most can be accommodated by a button addition, color change or text adjustment. For example, our findings showed that doctors participating in and driving research have a different view and level of reflection on technologies such as AI in the clinic. Some doctors in the hospital ward simply don't have the time to reflect differently on a tool just because it is AI, compared to a non-AI digital medical device. However, as the self-learning aspects of AI do pose new challenges This investigation bridges the gap between clinicians and AI designers through interactional expertise, emphasizing social responsibility and the importance of anthropology-driven design. Our creole VSD/VtM methodology combines participatory design approaches with anthropological methods to create socially responsible and user-centered technologies. By focusing on comprehensibility and usability, our design promotes better clinical decision-making and supports the integration of AI into healthcare practices, ultimately contributing to the improvement of patient care and the advancement of medical technology.

References

Agrawal, A., Gans, J., & Goldfarb, A. (2022). *Prediction Machines, Updated and Expanded: The Simple Economics of Artificial Intelligence*. Harvard Business Press. (Google-Books-ID: 8MBYEAAAQBAJ)

Andreasen, L. A., Tabor, A., Nørgaard, L. N., Taksøe-Vester, C. A., Krebs, L., Jørgensen, F. S., ... Tolsgaard, M. G. (2021). Why we succeed and fail in detecting fetal growth restriction: A population-based study. *Acta Obstetricia et Gynecologica Scandinavica*, *100*(5), 893–899. Retrieved 2024-02-29, from https://onlinelibrary.wiley.com/doi/abs/10.1111/aogs.14048 doi: 10.1111/aogs.14048

Anyoha, R. (2017, August). *The History of Artificial Intelligence.* Retrieved 2024-05-05, from https://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/

Armstrong, P. B. (2005). Phenomenology. In *The Johns Hopkins guide to literary theory and criticism* (Second ed ed.). Baltimore (Md.): the Johns Hopkins university press.

Bergmann, D. (2023, November). *What Is Reinforcement Learning From Human Feedback (RLHF)? | IBM.* Retrieved 2024-06-06, from https://www.ibm.com/topics/rlhf

Bohr, A., & Memarzadeh, K. (2020). The rise of artificial intelligence in healthcare applications. *Artificial Intelligence in Healthcare*, 25–60. Retrieved 2024-05-15, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7325854/ doi: 10.1016/B978-0-12-818438-7.00002-2

Borning, A., & Muller, M. (2012, May). Next steps for value sensitive design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1125–1134). Austin Texas USA: ACM. Retrieved 2024-05-06, from https://dl.acm.org/doi/10.1145/2207676.2208560 doi: 10.1145/2207676.2208560

Brinkmann, S. (2020). Kvalitet i kvalitative studier. In S. Brinkmann & L. Tanggaard (Eds.), *Kvalitative metoder* (pp. 657–670). København: Hans Reitzels Forlag. Retrieved 2024-06-06, from https://hansreitzel.dk/products/kvalitative-metoder-bog-20740-9788741277264#

Børsen, T. (2013). Identifying Interdisciplinary Core Competences in Techno-Anthropology: Interactional Expertise, Social Responsibility Competences, and Skills in Anthropology-Driven Design. In T. Børsen & L. Botin (Eds.), *What is Techno-Anthropology?* (pp. 35–66). Aalborg: Aalborg Universitetsforlag.

Collingridge, D. (1980). *The Social Control of Technology*. St. Martin's Press. (Google-Books-ID: hCSdAQAACAAJ)

Collins, H., & Evans, R. (2003). Constructing Expertise: In a Third Wave of Science Studies? *Social Studies of Science*, *33*(3), 419–434. Retrieved 2024-06-06, from https://www.jstor.org/stable/3183125 (Publisher: Sage Publications, Ltd.)

Corrales Compagnucci, M., Fenwick, M., Haapio, H., Minssen, T., & Vermeulen, E. P. (2020). Technology-Driven Disruption of Healthcare & 'UI Layer' Privacy-by-Design. *SSRN Electronic Journal*. Retrieved 2024-05-15, from https://www.ssrn.com/abstract=3611702 doi: 10.2139/ssrn.3611702

Dahlin, E. (2021, March). Mind the gap! On the future of AI research. *Humanities and Social Sciences Communications*, *8*(1), 1–4. Retrieved 2024-03-21, from https://www.nature.com/articles/s41599-021-00750-9 doi: 10.1057/s41599-021-00750-9

Dreyfus, S., & Dreyfus, H. (1980, February). A Five-Stage Model of the Mental Activities Involved in Directed Skill Acquisition. *Distribution*, 22.

European-Commission. (2012, October). *Charter of Fundamental Rights of the European Union.* Retrieved 2024-06-03, from http://data.europa.eu/eli/treaty/char_2012/oj/eng

European-Commission. (2019, April). *Ethics guidelines for trustworthy AI \textbar Shaping Europe's digital future.* Retrieved 2024-05-20, from https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Fischer, A., Rietveld, A., Teunissen, P., Hoogendoorn, M., & Bakker, P. (2023, October). What is the future of artificial intelligence in obstetrics? A qualitative study among healthcare professionals. *BMJ Open*, *13*(10), e076017. Retrieved 2024-06-06, from https://bmjopen.bmj.com/content/13/10/e076017 doi: 10.1136/bmjopen-2023-076017

Friedman, B., & Hendry, D. (2019). *Value Sensitive Design: Shaping Technology with Moral Imagination.* doi: 10.7551/mitpress/7585.001.0001

Friedman, B., Kahn, P., & Borning, A. (2003, January). Value Sensitive Design: Theory and Methods.

Friedman, B., & Kahn, P. H. (1992, January). Human agency and responsible computing: Implications for computer system design. *Journal of Systems and Software*, *17*(1), 7–14. Retrieved 2024-03-18, from https://linkinghub.elsevier.com/retrieve/pii/016412129290075U doi: 10.1016/0164-1212(92)90075-U

Friedman, B., & Nissenbaum, H. (1996, July). Bias in computer systems. *ACM Transactions on Information Systems*, *14*(3), 330–347. Retrieved 2024-05-06, from https://dl.acm.org/doi/10.1145/230538.230561 doi: 10.1145/230538.230561

Halkier, B. (2020). Fokusgrupper. In S. Brinkmann & L. Tanggaard (Eds.), *Kvalitative metoder* (pp. 167–184). København: Hans Reitzels Forlag. Retrieved 2024-06-06, from https://hansreitzel.dk/products/kvalitative-metoder-bog-20740-9788741277264

Hammersley, M., & Atkinson, P. (2019). *Ethnography: Principles in Practice.* doi: 10.4324/9781315146027

IBM. (n.d.-a). *What is Deep Learning? | IBM.* Retrieved 2024-06-06, from https://www.ibm.com/topics/deep-learning

IBM. (n.d.-b). *What Is Supervised Learning? | IBM.* Retrieved 2024-06-06, from https://www.ibm.com/topics/supervised-learning

IBM. (n.d.-c). *What Is Unsupervised Learning? | IBM.* Retrieved 2024-06-06, from https://www.ibm.com/topics/unsupervised-learning

Ihde, D. (1990). *Technology and the Lifeworld: From Garden to Earth.* Indiana University Press. Retrieved 2024-03-30, from https://philarchive.org/rec/IHDTAT-3

Jacob, P. (2023). *Intentionality* (Spring 2023 ed.). Metaphysics Research Lab, Stanford University. Retrieved 2024-03-18, from https://plato.stanford.edu/archives/spr2023/entries/intentionality/

Jensen, T. E. (2012). Intervention by Invitation: New Concerns and New Versions of the User in STS. *Science Studies*, *25*(1), 13–36.

Ke, G., Pan, L., Lie, G., Rocca, P., & Wu, J. (2021). 3,8 Engineering and Artificial Intelligence. In *Engineering for sustainable development: delivering on the Sustainable Development Goals - UNESCO Digital Library* (pp. 106–110). UNESCO. Retrieved 2024-06-06, from https://unesdoc.unesco.org/ark:/48223/pf0000375644

Kudina, O., & Verbeek, P.-P. (2019, March). Ethics from Within: Google Glass, the Collingridge Dilemma, and the Mediated Value of Privacy. *Science, Technology, & Human Values*, *44*(2), 291–314. Retrieved 2024-03-24, from https://doi.org/10.1177/0162243918793711 (Publisher: SAGE Publications Inc) doi: 10.1177/0162243918793711

Lieberman, H. (2009). User Interface Goals, AI Opportunities. *AI Magazine*, *30*(4), 16–22. Retrieved 2024-05-15, from https://www.proquest.com/docview/208118941/abstract/8877C8C9F54B4280PQ/1 doi: 10.1609/aimag.v30i4.2266

Lin, M., Feragen, A., Bashir, Z., Tolsgaard, M. G., & Christensen, A. N. (2022, November). *I saw, I conceived, I concluded: Progressive Concepts as Bottlenecks.* arXiv. Retrieved 2024-05-22, from http://arxiv.org/abs/2211.10630

Meske, C., & Bunde, E. (2023, April). Design Principles for User Interfaces in AI-Based Decision Support Systems: The Case of Explainable Hate Speech Detection. *Information Systems Frontiers*, *25*(2), 743–773. Retrieved 2024-05-15, from https://doi.org/10.1007/s10796-021-10234-5 doi: 10.1007/s10796-021-10234-5

Murphy, M. (2018, August). *10 Tips for Designing Clinical Device Interfaces | Frontend.com.* Retrieved 2024-06-02, from https://www.frontend.com/thinking/10-tips-for-designing-clinical-device-interfaces/ (Section: Thinking)

OpenAI. (n.d.). *Preserving languages for the future.* Retrieved 2024-06-06, from https://openai.com/index/government-of-iceland/

Radboud Reflects. (2018, January). *How Technology Changes Us | Lecture with Don Ihde and Peter-Paul Verbeek.* Retrieved 2024-05-17, from https://www.youtube.com/watch?v=hmBgJjfjG7Q

Reverberi, C., Rigon, T., Solari, A., Hassan, C., Cherubini, P., & Cherubini, A. (2022, September). Experimental evidence of effective human–AI collaboration in medical decision-making. *Scientific Reports*, *12*(1), 14952. Retrieved 2024-06-06, from https://www.nature.com/articles/s41598-022-18751-2 doi: 10.1038/s41598-022-18751-2

Roessler, B., & Mokrosinska, D. (2013). *Privacy and social interaction.* Retrieved 2024-06-06, from https://journals.sagepub.com/doi/full/10.1177/0191453713494968

Roller, M. R., & Lavrakas, P. J. (2015). *Applied qualitative research design: A total quality framework approach.* New York, NY, US: The Guilford Press. (Pages: xviii, 398)

Salomon, L. J., Alfirevic, Z., Da Silva Costa, F., Deter, R. L., Figueras, F., Ghi, T., ... Yeo, G. (2019, June). ISUOG Practice Guidelines: ultrasound assessment of fetal biometry and growth. *Ultrasound in Obstetrics & Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology*, *53*(6), 715–723. doi: 10.1002/uog.20272

Searle, J. R. (1980, September). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–424. Retrieved 2024-06-06, from https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/abs/minds-brains-and-programs/DC644B47A4299C637C89772FACC2706A doi: 10.1017/S0140525X00005756

Sheikh, H., Prins, C., & Schrijvers, E. (2023). Artificial Intelligence: Definition and Background. In H. Sheikh, C. Prins, & E. Schrijvers (Eds.), *Mission AI: The New System Technology* (pp. 15–41). Cham: Springer International Publishing. Retrieved 2024-05-05, from https://doi.org/10.1007/978-3-031-21448-6_2 doi: 10.1007/978-3-031-21448-6_2

Smits, M., Bredie, B., Goor, H. v., & Verbeek, P.-P. (2019, November). Values that Matter: Mediation theory and Design for Values. In *Research Perspectives in the era of transformations: Conference proceedings* (pp. 396–407). Academy for Design Innovation Management. Retrieved 2024-03-18, from https://research.utwente.nl/en/publications/values-that-matter-mediation-theory-and-design-for-values doi: 10.33114/adim.2019.03_203

Smits, M., Ludden, G., Peters, R., Bredie, S. J. H., van Goor, H., & Verbeek, P.-P. (2022, January). Values that Matter: A New Method to Design and Assess Moral Mediation of Technology. *Design Issues*, *38*(1), 39–54. Retrieved 2024-03-18, from https://doi.org/10.1162/desi_a_00669 doi: 10.1162/desi_a_00669

Spradley, J. P. (1979). *The ethnographic interview / james p. spradley.* New York: Holt, Rinehart and Winston.

Sundhedsstyrelsen. (2023). *Specialeplan for gynækologi og obstetrik.* Retrieved 2024-06-06, from http://www.sst.dk/da/Fagperson/Sundhedsvaesenets-rammer-og-uddannelser/Specialeplaner-for-sygehusvaesnet/Gaeldende-specialeplan/Specialeplan-for-gynaekologi-og-obstetrik

Swain, J. (2018). *A hybrid approach to thematic analysis in qualitative research: Using a practical example.* doi: 10.4135/9781526435477

Tanggaard, L., & Brinkmann, S. (2020). Interviewet: Samtalen som forskningsmetode. In S. Brinkmann & L. Tanggaard (Eds.), *Kvalitative metoder* (pp. 33–64). København: Hans Reitzels Forlag. Retrieved 2024-06-06, from https://hansreitzel.dk/products/kvalitative-metoder-bog-20740-9788741277264

Tariq, R. A., & Sharma, S. (2024). Inappropriate Medical Abbreviations. In *StatPearls.* Treasure Island (FL): StatPearls Publishing. Retrieved 2024-04-07, from http://www.ncbi.nlm.nih.gov/books/NBK519006/

Tolsgaard, M. G., Boscardin, C. K., Park, Y. S., Cuddy, M. M., & Sebok-Syer, S. S. (2020, December). The role of data science and machine learning in Health Professions Education: practical applications, theoretical contributions, and epistemic beliefs. *Advances in Health Sciences Education*, *25*(5), 1057–1086. Retrieved 2024-02-29, from https://doi.org/10.1007/s10459-020-10009-8 doi: 10.1007/s10459-020-10009-8

Topol, E. J. (2019, January). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, *25*(1), 44–56. Retrieved 2024-02-29, from https://www.nature.com/articles/s41591-018-0300-7 doi: 10.1038/s41591-018-0300-7

Ultrasound. (2023). Retrieved 2024-06-06, from https://www.nibib.nih.gov/science-education/science-topics/ultrasound

Umbrello, S., & van de Poel, I. (2021). Mapping Value Sensitive Design onto AI for Social Good Principles. *AI and Ethics*, *1*(3), 283–296.

United-Nations. (2012). *Goal 3 \textbar Department of Economic and Social Affairs.* Retrieved 2024-06-02, from https://sdgs.un.org/goals/goal3

van den Hoven, J., Vermaas, P. E., & van de Poel, I. (2015). Design for Values: An Introduction. In J. van den Hoven, P. E. Vermaas, & I. van de Poel (Eds.), *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains* (pp. 1–7). Dordrecht: Springer Netherlands. Retrieved 2024-05-06, from https://doi.org/10.1007/978-94-007-6970-0_40 doi: 10.1007/978-94-007-6970-0_40

van de Poel, I. (2013). Translating Values into Design Requirements. In D. P. Michelfelder, N. McCarthy, & D. E. Goldberg (Eds.), *Philosophy and Engineering: Reflections on Practice, Principles and Process* (pp. 253–266). Dordrecht: Springer Netherlands. Retrieved 2024-03-30, from https://doi.org/10.1007/978-94-007-7762-0_20 doi: 10.1007/978-94-007-7762-0_20 van de Poel, I., & Royakkers, L. (2011, March). The Responsibilities of Engineers. In *Ethics, Technology, and Engineering: An Introduction* (pp. 6–30). John Wiley & Sons. (Google-Books-ID: XHNxT1wikPEC)

Verbeek, P.-P. (2008, March). Obstetric Ultrasound and the Technological Mediation of Morality: A Postphenomenological Analysis. *Human Studies*, *31*(1), 11–26. Retrieved 2024-05-20, from https://doi.org/10.1007/s10746-007-9079-0 doi: 10.1007/s10746-007-9079-0

Designing Responsible AI Interfaces in Obstetric Ultrasound

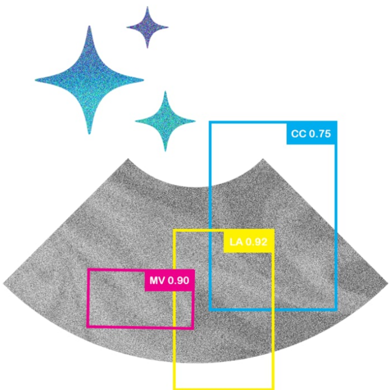
Part 2: Article

Integrating Values into AI-UI Design: *Responsible UI Design for an AI for Estimation of Fetal Weight*

Andreas Biering von Essen

&

Andreas Scott Magle Pedersen



Supervisor:

Anders Hansen Henten

Noname manuscript No.
(will be inserted by the editor)

Integrating Values into AI-UI Design

Responsible User Interface Design of an AI for Estimation of Fetal Weight

Andreas Biering von Essen · Andreas Scott Magle Pedersen · Chun Kit Wong · Martin Grønnebæk Tolsgaard · Mary Ngo

Received: date / Accepted: date

Abstract Artificial Intelligence (AI) has the potential to bring change to the field of obstetrics. In this paper we explore the development of a User Interface (UI) an AI model aimed at assessing ultrasound images and providing decision support of estimating fetal wheigt (EFW). Interactive human-AI collaboration could address diagnostic uncertainties and educational gaps among clinicians performing obstetric ultrasound procedures. However, the successful integration of AI into a clinical setting requires careful consideration of UI design to ensure a responsible human-AI collaboration. Drawing on principles of Value Sensitive Design (VSD), particularly the Values that Matter (VtM) approach, this study investigates how the needs of healthcare practitioners engaged in fetal weight estimation using ultrasound can be translated into design requirements for AI-driven obstetric ultrasound. The methodology combines qualitative research methods, including semi-structured interviews and user testing sessions with diverse stakeholders, to inform the design process and UI-development. By aligning AI-development with societal values and stakeholder needs, this study aims to foster responsible innovation and promote the effective utilization of AI in clinical settings.

Keywords Fetal Growth Assessment · Obstetric Care · Ultrasound imaging · AI in Healthcare · User Interface in AI · Value Sensitive Design

Andreas Scott Magle Pedersen
Department of Planning, Aalborg University , Copenhagen Denmark
Tel.: +45 31322514
E-mail: asmp19@student.aau.dk

Andreas Biering von Essen
Department of Planning, Aalborg University , Copenhagen Denmark
Tel.: +45 30573550
E-mail: aessen22@student.aau.dk

1 Introduction

Artificial intelligence (AI) that can aid in acquisition and interpretation of images during image-based examinations is steadily gaining ground in healthcare settings Bohr & Memarzadeh (2020) This includes AI-assisted predictions in ultrasound imaging, where AI has the potential to enhance the detection rates of pregnancies at risk by providing clinicians with feedback that can support the clinical decision making (Tolsgaard et al., 2020; Topol, 2019; Bohr & Memarzadeh, 2020; Fiorentino et al., 2023; Yi et al., 2021). Interactive human-AI-collaboration can support clinical decision making and improve the quality of diagnostics without attempting to replace the clinicians (Bohr & Memarzadeh, 2020; Dégallier-Rochat et al., 2022).

This paper examines an AI model in development which is capable of assessing the quality of ultrasound images and can potentially provide decision support for clinicians estimating fetal weight (EFW). We work within the scope of the SONAI-project, an interdisciplinary research group developing AI-feedback for obstetric ultrasound at the Danish Pioneer Center for AI. The support from AI can limit disparities and enhance the quality of obstetric care by addressing diagnostic uncertainty and educating clinicians performing obstetric ultrasound procedures (Topol, 2019; Bohr & Memarzadeh, 2020).

While AI has the potential to improve quality of care, it also has the capability to influence clinical practice and decision-making (Mintz & Brodie, 2019). This calls for an investigation of how clinicians engage with AI-feedback and how it is presented, interpreted and utilized by the clinician. The user interface (UI) plays a significant role in defining how the AI is actually used and perceived. The UI connects the black boxed inner workings of the AI with actionable information (Meske & Bunde, 2023; Lieberman, 2009). Effective UI design can increase comprehension for the users and help to address ethical considerations or biases promoting a more responsible human-AI-collaboration (Corrales Compagnucci et al., 2020). A poor UI can consequently result in a negative user experience and may lead to the rejection of the very utility of the AI by the end user (Lieberman, 2009). We aim to investigate the integration of AI into clinical practice via responsible design of UI. In this study we employ social scientific methods and perform qualitative investigations involving clinicians with diverse levels of clinical expertise to examine the user-technology dynamics at play. The objective is to derive UI design requirements for the AI developed by the SONAI-project for obstetric ultrasound. This is to ensure that its development and deployment in the clinic aligns with societal values and fosters responsible innovation.

1.1 The need in the Clinic

Fetal ultrasound plays a key role in the screening for, and the management of, fetal growth abnormalities (Salomon et al., 2019). Growth scans enable healthcare professionals to identify conditions that are associated with an increased risk of adverse maternal-fetal outcomes (L. Andreasen et al., 2021). Early and accurate identification of growth restrictions is crucial for the timely planning of interventions and is identified as an essential factor for ensuring favorable perinatal outcomes and to reduce the incidence of birth injuries and complications (L. A. Andreasen et al., 2021). Following the International Society of Ultrasound in Obstetrics and Gynecology (ISOUG) Practice Guidelines (2019) to perform the growth scans to EFW, at least three fetal biometric parameters or standard planes are needed: head circumference, abdominal circumference and the femur diaphysis length.

Many factors and specific skills related to acquisition and interpretation of scans have an impact on EFW and can complicate the examination. To accurately perform EFW, it is vital that the operator is proficient in skills ranging from image optimization and interpretation to probe manipulation (Nicholls et al., 2014). External factors like high maternal BMI or the fetal representation during ultrasound scans can increase the difficulty of the procedure (Nicholls et al., 2014; Paladini, 2009). In ambiguous or challenging cases the operator must decide whether the assessment requires follow-up reviews or re-assessments by another clinician (Govaerts et al., 2011). However, even the act of identifying when a case is ambiguous or challenging can be unclear to an independently learning clinician, and the feedback from expert clinicians is costly and limited due to the demands of their clinical departments (Govaerts et al., 2011). By leveraging advanced algorithms and machine learning techniques, SONAI aims to provide real-time support to clinicians during EFW procedures. An AI decision support system could help to navigate the complexities posed by ambiguous cases and augment the clinical decision making with actionable feedback, thereby enhancing accuracy in fetal weight estimation (Dégallier-Rochat et al., 2022; Topol, 2019)

If a technology is early in development, plenty of the design decisions can be changed, but the design team does not necessarily have sufficient knowledge about how it will affect the end users and the broader society (Kudina & Verbeek, 2019). This especially applies for AI-technologies where some features or uses are neither foreseen or intended by design (Umbrello & van de Poel, 2021). Technical and social issues tend to be considered analytically distinct from the specific technical features in research and development of AI (Dahlin, 2021). This separation can be contradictory since technology mediates values, co-shaping behaviors and experiences of the users of a technology (Dahlin, 2021). By conducting qualitative research within the operational environment between the user and the technology, it is possible to examine the dynamics, evaluate ethical considerations and societal values (Dahlin, 2021). These important insights into what is required for development and deployment of an AI-system can provide a basis for a UI that aligns with the needs of the clinic.

1.2 Research question

Further exploration of the representation of the UI and societal implications of AI-assisted feedback when introduced in a clinical setting is needed to ensure feasible and sustainable AI systems that can benefit in the clinic. We will address this relating to the SONAI-project with the following research question:

How can the needs of healthcare practitioners engaged in fetal weight estimation using ultrasound be translated into design requirements for the development of user interfaces for AI-driven obstetric ultrasound by employing principles of Value Sensitive Design?

In the subsequent sections we will explain Value Sensitive design (VSD) and how we have employed the methodological approach of Values that Matter (VtM) in order to identify values and take them into account in design practices for a responsible UI in AI development. We will detail the process of defining design requirements through empirical and contextual investigations of the stakeholders, which can be employed in the outset of the technological design process, when developing AI for EFW.

2 Values in Design

Since the UI of an AI product taps into its inner workings and mediates the user experience, many of the same AI-specific theoretical and ethical concerns apply to a study of AI-UI. In order to translate our empirical investigations into UI design requirements, we use the theoretical framework and methodologies of value sensitive design (VSD).

2.1 Value Sensitive Design

In the methodology of VSD, technology can be developed through three iterative phases (Friedman et al., 2003): First *the conceptual investigations*, in which values are identified and ordered, informed by philosophy and by the stakeholders themselves. Second *the empirical investigations*, where the values are evaluated in the design context using methods from social science. Third *the technical investigations*, where existing and potential technologies are investigated both retrospectively and proactively, followed by the design of a novel prototype or product. Whether used explicitly or implicitly in a given analysis, a fundamental assumption of VSD is that the intersecting results of these investigations will yield a product that is morally compatible with society.
Defining values is an important part of the VSD literature. Values were defined by Friedman as "*what is important to people in their lives, with a focus on ethics and morality*" (Friedman & Hendry, 2019, p. 24). This human-centered approach to values is also recognizable in van de Poel and Royakkers' definition of values as "*... lasting convictions or matters that people feel should be strived for in general ...*" (van de Poel, 2013, p. 27). Friedman operates from a set list of values. Although there are obvious shortcomings of a heuristic list of values (such as privileging values that benefit specific stakeholder groups) Friedman deems it more productive to use a fixed set of values to more easily build upon previous work (Friedman & Hendry, 2019). At the other end of the spectrum Smits et al. (2019, 2022) see values as arising when humans engage with technology and thus ought to be identified empirically by engaging with stakeholders in situ. Specifically for VSD in AI-development Umbrello & van de Poel (2021) suggest a mixed approach of value identification, both informed by normative dimensions and derived from the specific context. This is because there is a perceived consensus that moral issues surrounding AI differ from traditional issues surrounding digital and information technologies. Some sort of value list for typical AI issues can prove productive, but it is also important to elicit values from the bottom and up (Umbrello & van de Poel, 2021). They thus provide three sources for human values:

- **Values promoted by the design:** These values are an orientation towards contributing to the betterment of society, such as contributing to the UN's Sustainable Development Goals (SDG's).
- **Values respected by the design:** By embodying principles of certain values into the design. These values could be sourced from an EU high-level expert group on AI and are focused on avoiding societal harm.
- **Context specific values:** These are the values that are identified in the context, not covered by the other two sources.

2.2 Technological Mediation

Technological mediation offers a nuanced approach to addressing the interaction between humans, technology and their environment. It is attributed to the Dutch philosopher Peter-Paul Verbeek, building on the post-phenomenological works of the American philosopher Don Ihde (1990). It posits that technology isn't merely external to humans, but something that mediates the world, and changes our way of interacting with the world and each other. One example is how digital technologies didn't just raise privacy questions but have fundamentally redefined the concept of privacy itself from something spacious and tangible to a question of data ownership (Kudina & Verbeek, 2019).

2.3 Values that Matter

Technological mediation has been methodized into an adaptation of VSD, namely the Values that Matter (VtM) approach (Smits et al., 2019, 2022). By using technological mediation VtM provides a framework for understanding how values themselves are transformed by new technologies. This can be used in the process of responsible design of technology, by allowing designers to anticipate technologically induced value mediation in an informed way. The authors operationalize this by introducing their own tripartite methodology, focused on finding the "*right balance between speculation and experimentation*" (Smits et al., 2019, p. 43).

Although methodologically adjusted, the three phases of VtM are based on the conceptual, empirical, and technological investigations of VSD. The first phase, *explore*, involves "*mapping out the context of the design problem*" (Smits et al., 2019, p. 399), which includes identifying important actors and developing a value framework. This phase combines conceptual and empirical investigations. In the *conceptualizing phase*, a prototype is created based on the defined groups and values, incorporating the technical investigations. During the *anticipation phase*, the technology is measured at the threshold of society. According to the authors, technology fundamentally redefines how humans engage with values. This can be explored in practice by introducing the concept to the relevant actors in an authentic context. Thus it can be examined how a technology mediates the identified values differently than in a context without the technology (or with a previous iteration of the technology). These value experiences can be identified and translated into design requirements and make the basis for a *re-conceptualization*.

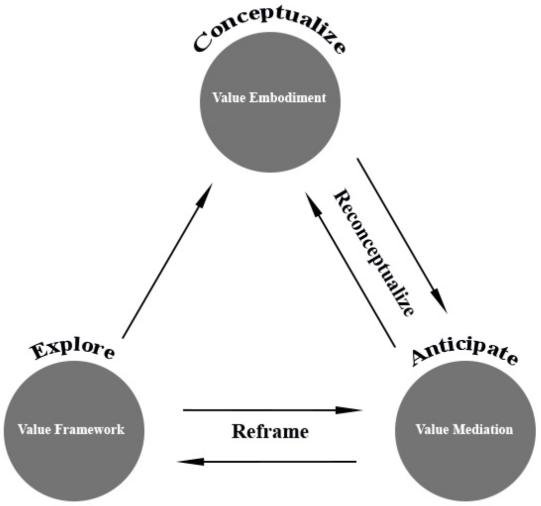


Fig. 1 Values that Matter model (Smits et al., 2022)

In the following section we will detail on the identification of a value framework, bringing together value-based design and value mediation of the technology.

3 Method

To explore the needs of the clinicians, we employed qualitative research methods with a multifaceted strategy for data collection. The primary method of data collection was semi-structured interviews – this includes focus group interviews with presentations of UI mock-ups as well as user testing sessions with AI-prototype iterations. The sampling method for participants was purposely selected to include physicians specializing in gynecology and obstetrics as well as physicians with little to no experience in ultrasound. A total number of six interviews were conducted, as well as one larger focus group. The participants included three 8th semester medical students, two entry position MD's with a recent Ph.D. in gynecology and obstetrics, two unspecialized MD's, a senior obstetrician and a pregnant individual in the third trimester. The focus group consisted of twelve doctors with varying levels of expertise, ranging from trainee doctors within gynecology and obstetrics to senior clinicians.
The participants were presented with a working prototype (Fig. 2), and AI-user interfaces ranging from minimal feedback to more advanced interfaces, containing elaborate descriptions and an increased number of UI elements. The interviews were performed in extension of a user testing scenario. This was done in either a realistic clinical setting where AI was introduced as the participants performed the EFW on a pregnant volunteer, or an artificial setting where the participants tried the AI on a medical phantom mimicking a pregnant individual.

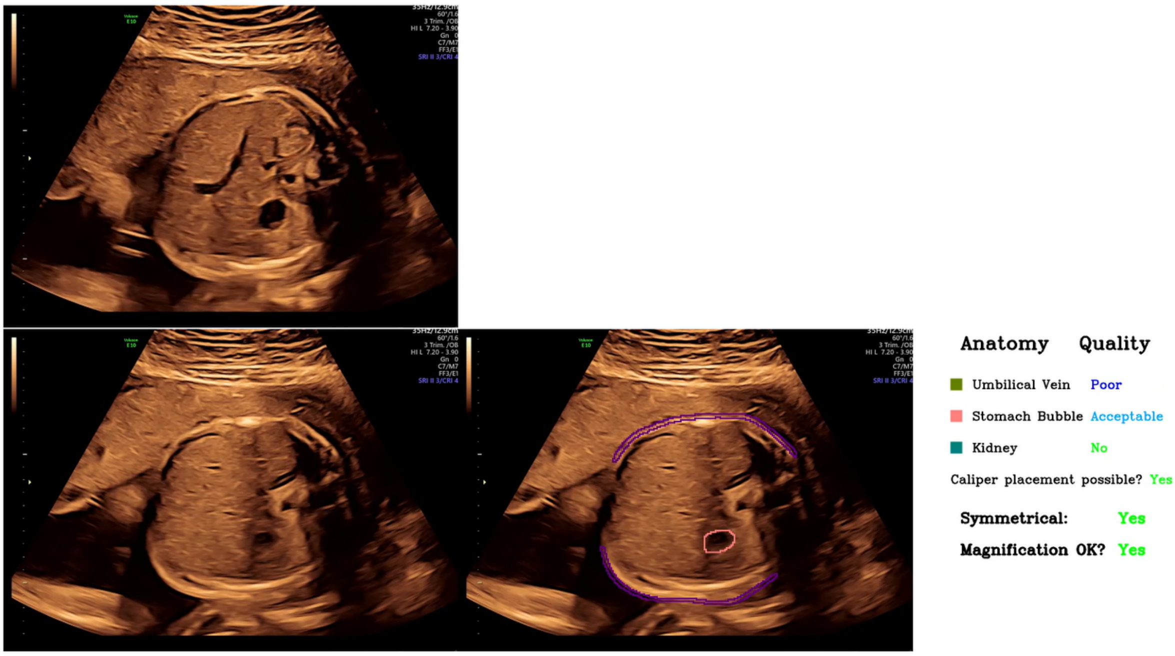


Fig. 2 First iteration of the AI User Interface

The focus group was presented with footage of the UI of a functioning AI prototype and UI mock-ups and asked to evaluate and reflect on their experiences, needs and requirements in a general sense. All the interviews were transcribed, pseudonymized and coded using NVIVO for further analysis. We obtained informed consent from each participant in full confidentiality.

4 Value-Centric Analysis: Framework Definition and Synthesis

In the following section we will present and analyze the collected data with VSD and VtM as the theoretical and methodological frameworks. Based on the analyses the design requirements will be summarized in a table that serves as the outset for the re-conceptualization of a novel UI-design.

4.1 Explore

The first step when employing the VtM approach is the theorization of a value framework. Friedman's adherence to a set list of 12 values is explained in part as to allow to build on previous work, using said list (Friedman & Hendry, 2019). Although we do not build on Friedman's value list, we do see the relevance in being able to build upon previous work by re-using value frameworks. In our case, we build on Umbrello & van de Poel's (2021) value framework for AI-specific values supplemented with context-specific values.

4.1.1 Values promoted by the design

Overall, the SONAI-project aims to improve the quality of care by bringing AI to the medical field of obstetrics such as AI feedback for EFW. The third SDG (Good health and well-being) and especially the Targets 3.1, 3.2, 3.8, and 3.c relating to maternal and prenatal care as well as training of healthcare workers (United Nations, 2012), have potential to be improved by AI-solutions

4.1.2 Values Respected by the Design

As suggested by Umbrello & van de Poel (2021), we source a list of values based on the ethical principles by the high level EU expert group on AI (AI HLEG). These four ethical principles are based on the EU Charter of Fundamental Rights of the (European Commission, 2012) and respecting these principles ensures normative compatibility within the EU ahead of regulative compatibility (European Commission, 2019). The AI HLEG reasons these principles as such:

- **Human Autonomy:** AI systems should be designed to augment and empower humans, not take over decision making power.
- **Prevention of Harm:** AI systems should not increase power gaps in various social configurations, and foreseeable scenarios of misuse must be accommodated before implementation.
- **Fairness:** The costs and benefits of AI systems should be distributed. Likewise bias in training data should not unjustifiably impair freedom of choice. An analysis by an AI should be contestable and not taken for granted.
- **Explicability:** AI systems are as transparent as possible. The capabilities, limitations and purposes are openly communicated.

4.1.3 Context specific values

The participants' experiences with the mock-ups and prototype interfaces informed by the interviews constitute the method for value identification of the context specific values. This approach facilitated a dynamic exchange of feedback and allowed us to capture the participants' reactions and preferences as they interacted with the prototypes.

We distinguish between the values of *providing good care* and *receiving good care* in order to emphasize where the primary impacts of the design requirements would be. In practice these are intertwined, and we sourced both sets of values and norms from clinicians. *Providing good care* was a central value and considered important by our participants as well as the norms that constitute the accomplishment of this value. This includes an adequate workflow with timely and equitable interventions centered around the patient. We observed that it was important that the AI diffused uncertainty in the clinic and augmented the clinician's conviction regarding accuracy, reliability and effectiveness of medical decisions made within a clinical context. This is also reflected in the professional integrity and accountability between the patient and the clinicians fostering trust and transparency in clinical decisions. Lastly, we refer to the features of the AI-UI that fosters independent learning under the *emancipation*-value.

4.2 Conceptualize

At the beginning of our investigations, a prototype with a working UI had already been developed by the design team (Fig. 2). However this prototype was not explicitly designed with a value framework as the basis of development and thus still needs translation of values into design requirements for further development. To investigate how the AI mediates the users actions, perceptions, and values we bring this prototype to the users. The UI of the prototype comes in two versions – minimum feedback and maximum feedback. The following sections aim to anticipate how these two versions transform value experiences, in an effort to re-conceptualize the UI elements' mediation of relevant values relating to care and AI.

4.3 Anticipate

The results of our investigations are listed in Table 1. The positive value experience column shows where the device holds positive mediation of a value, and the negative value experience column shows the opposite. For the purposes of illustrating the potentials of the device, we deviate from the VTM framework by also including conceivable value experiences that were not empirically found. This is due in part to our inclusion of a set list of values placed above the context specific values, and to make it more clear what a given design recommendation aims to accommodate or avoid.

4.3.1 SDG 3

According to Umbrello & van de Poel (2021) it is important that an AI explicitly contributes to the betterment of society besides merely avoiding causing harm. For the AI to reach a state where it increases maternal and newborn health in general, it needs to make intuitive sense to the operator. The contribution to early detection rates of fetal abnormalities could mitigate the complications of pregnancies at risk. In general, even participants who had no prior introduction to the procedure and what to look for had a degree of success with both the minimum and maximum feedback version of the AI. We see this as an indication that even in its current conception, the AI does show potential for unsupervised learning and the inclusion of unspecialized clinical staff in a highly specialized environment. However it is still difficult for novice ultrasound operators to manipulate the ultrasound probe into a standard plane, indicating a need for a guidance system. Relating to the third SDG, the technology thus shows a potential for increasing societal good and improving maternal and newborn health outcomes by bringing AI feedback to the field of obstetrics.

Table 1 Value Mediation of Users and Potential Users Interacting with the UI

Value	Design Norm	Positive value experiences	Negative value experiences	Design recommendation
SDG 3 Good health and well-being	Target 3.1 Reduce maternal mortality	The technical and professional abilities in the healthcare setting enables identifying and responding appropriately to a EFW	Preventable complications are not accommodated. Unnecessary intrusive interventions are concluded based on the AI-feedback.	UI that helps with identification of standard planes when performing EFW. The AI can help with clearly identifying anatomy that indicates a wrong standard plane as well as a correct standard plane.
	Target 3.2 Prevent death of newborns and children under 5			
	Target 3.8 Achieve universal health coverage	The device make sense in theory, training and practice	The practices surrounding / required by the device doesn't match the context of the clinic.	Accessible and understandable UI with instructions that are intuitive to an unspecialized healthcare professional
	Target 3.c Secure training of health workforce			
Human Autonomy	AI should not take over too much decision- making power	The device augments the practitioner's decision making when taking optimal pictures for accurate EFW.	The practitioner 'games' the scan, looking for all parameters to 'turn green'.	- The AI does not output an actionable analysis based on the scan. - Help to flag ambiguous cases where expert assistance is needed. - An indication of how sure the AI is of having found the correct SP e.g. a per-centage indicator to avoid prejudice bias.
Prevention of harm	Protection against statistical prejudice			
Fairness	There should be a human who is able to be held accountable for the AI	- Accountability is traceable and on site. - The reasoning behind AI-interpretations are transparent.	- Accountability and responsibility is symbolic - the AI's analyses are 'taken for granted as true'. - The AI-reasoning is considered opaque.	- The AI's maximum information layout should provide a rationally sound deductible reasoning for its estimations, that can be intellectually contested by a human operator. - By default the 'maximum information' layout is turned on.
Explicability	AI decisions should be contestable and rationally deductive			
Provide good care	Indisruptably support the clinical workflow	- The AI-interpretation is con-sidered as accurate - Minimal interruptions of clinical workflow.	Inaccurate handling of the US machine when the AI can't keep up with real-time	- Provide a visual indication of the current latency of the AI
	Enhance Clinical Certainty	The device improves confidence in decision-making.	The device makes false positives	An indication of how sure the AI is of having found the correct SP e.g. a percentage indicator.
Receive good care	Strengthen Integrity and Accountability	The patient can trace and understand the responsibilities	Fear that the AI makes an analysis that is 'mediated' by the Doctor. No human eyes on the test results.	The UI should show the logic of an analysis, and the logic should clearly correspond to medical literature.
	Support Patient Engagement	The patient experiences that their input is used and respected.	Worry that the AI interprets on the health of the baby.	Misinterpretable UI elements should be limited or changed.
	Increase Trust and transparency	- The patient can trust the clinical decisions based on interpretations that include AI in diagnostics.	The device is too esoteric and increases the distance between patient and practitioner.	- The UI should also be at the same 'height' as the pregnant person. - Interpretations of UI-elements from the patient should not lead to distrust in decisions from the clinicians.
Emancipation	Facilitate Independent learning	The development of skills can be controlled and managed by the clinician on their own.	The practitioner gets very good at performing scans with the AI, but becomes over-reliant on it.	- Toggle-able screen elements for practicing without "safety wheels" - Directional support system when performing EFW

4.3.2 AI HLEG

The AI augments the practitioner's decision making by providing support for finding the correct standard planes to be used in EFW assessments. However the AI also has a mediating impact on *human autonomy* by affecting the decision making to some extent even dictating the movements of the clinicians. During user-testing some of the participants would solely focus on making the UI turn 'green' and not so much on whether the ultrasound pictures were accurate enough to be used for EFW. In this way the operator 'games' the scans looking for 'all green'.
According to HLEG - AI systems should aim to be as transparent as possible (European Commission, 2019). The capabilities, limitations and purposes are openly communicated and to an extent explicable. One of the participants expressed that if the device is present in their department, then "*I have to trust that it works* (...) *and it must be approved and validated.*" (P5) The doctor in this case expressed a trust towards the approving bodies and regulation in place to ensure that medical devices only get implemented if they are tested and approved for clinical use. This notion is interesting when held up against the minimum feedback version of the UI relating to the values of *human autonomy* and *explicability.* If the information and reasoning behind the AI's analysis is hidden behind a green color, the responsibility is diffused. It becomes opaque as it is shared between the clinician performing the procedure, the doctor in charge of the department, the approval board that approved the device, and even the designers of the technology.

4.3.3 Care

Performing EFW requires a significant level of multitasking; the operator needs to both control an ultrasound machine, handle the probe, interpret the picture, care for the patient and now relate to information from an AI. The participants preferred different levels of feedback ranging from minimum to maximum. A place where the technology had a mediating impact was through cognitive load affecting the clinical workflow. Additional screen elements and variables that need to be checked could potentially over-complicate the procedure, counter-intuitively taking up even more of the clinician's mental capacity than without the AI support system. The clinicians also had to learn how to operate the AI while performing EFW and understand how to interpret the AI prediction:

> *"Machine learning and such... I am totally overloaded already with my own stuff. I just want* [a tool] *that is easy, good and functions – like all the other tools."* (P5)

Although training and education is associated with the use of a new device in the clinic, in an information-heavy environment, it would be problematic if every user had to assess whether or not a given tool is AI-trained or not, when using it for clinical decision making. One participant expressed an expectation that you should trust the analysis of medical devices even without understanding its technical inner workings:

> *"Yes, 100 procent I am going to trust the AI-interpretations."* (P5)

The participants in general also expressed initial enhanced confidence and elevated clinical certainty when performing EFW with AI-feedback. The real-time feedback provided was perceived as trustworthy for clinical decision making and provided reassurance to the operators that they were in the correct plane. However the clinical certainty was sometimes dictated by the results of the AI and were perceived as adequate for EFW. One observation was how a participant would uncritically freeze images based on feedback from the AI:

"I think it's a bit hard, since I don't really have any background information on this, like, I have to be confident in myself, and then I need approval from the AI tool (...) with the AI I would feel more confident, but right now I'm not confident at all! And then it says, you're probably right, and then I'm like, okay this is as good as it gets. . . " (P6)

If we relate this to the four values sourced from the HLEG, a minimum feedback version is in a direct value clash with all four Values. The ultrasound images required to perform EFW need to be explicable, confirming exactly why a certain plane is adequate for clinical decision making. By not displaying in the UI why the AI claims that the picture is good enough for an accurate weight estimation, the AI appears both deceptive, authoritative and insensitive towards its own bias if used in a situation it is not trained for. It is important that the AI augments the practitioner's decision-making by providing more context and strengthening certainty, and although a clinically approved tool can reasonably be expected to enhance the procedure, it does provide accountability issues if its analysis in hindsight led to the wrong intervention.

"I just researched around and tried to find something. And maybe there was a little bit... It was lagging." (P6)

To be able to provide good care, the operator needs to provide clinical decisions accurately and preferably in an adequate amount of time. Some participants were dependent on the feedback from the AI, even though the feedback was in real time, the exact speed did not match that of the ultrasound machine, with a slight delay. An EFW-procedure is more than just an instrumental procedure. It is a very personal situation shared between the patient and the clinician. Therefore, the dimension of the conversation between the ultrasound operator and the patient should also be acknowledged in the design requirements. Even though we only included one pregnant individual in this study, there was an interest in this introduction of the AI:

"If you come in for a scan, you want all the information you can possibly get." (P4)

Therefore it is also important that the UI does not cause unnecessary concern to a patient looking at the screen if the UI appears dangerous e.g. with a lot of red colors and warning-triangles. The patient did not mind the prototype UI however, and expressed that it was obvious that it related to the clinician and not the health of the fetus:

"I have not seen it as something where it said 'something is wrong' here. For me it was obvious that the feedback from the AI was related to the clinician." (P4)

When AI is introduced as a novel technology in the clinic when performing EFW it should be on a level that is not esoteric, increasing the distance between the patient and the clinician. The UI elements presented to the patient should therefore be explicable to the patients or their relatives as non-clinicians and should not be up for misinterpretations that could lead to distrust.

4.3.4 Emancipation

A place where the AI especially had a mediating impact was through emancipation of young clinicians supporting their learning of how to perform EFW. Some young doctors saw a learning potential in the AI, where they could locate standard planes on their own. The UI elements should allow a novice to get a sense of their performance, whether they are finding standard planes or not, even without a senior clinician providing them with feedback.

"Finally we can contribute as young doctors, the old doctors doesn't have the time to wait for my unsteady hands" (P1)

By utilizing the potential learning value the AI could support training of a healthcare workforce that supports the third SDG promoting well-being for all ages and especially the targets surrounding maternal and newborn health.

5 Results: Re-conceptualize

We conducted iterative design sessions, incorporating the feedback from our participants in order to produce low-fidelity mock-ups to re-conceptualize a better, more responsible UI. Changes were observed in almost all the participants' value experiences, and Smits et al. (2022) suggest that design requirements should be created if the concept holds negative mediation. However since we also sourced values from the SDGs and the AI HLEG some contradictions occurred where a design requirement accommodating the value experiences of the users would not respect these externally sourced values. In these instances, we opted to find a middle-ground, albeit favoring these higher-level values. The UI elements are selected based on data from the anticipation phase of the analysis and informed by Table 1 identifying design elements for optimizing value mediation. Using this table, we generated the design concept, which incorporated all requirements for optimal mediation of patients' values.

5.1 Universability

As the AI-solution in development has to work in an environment with non-specialist medical professionals such as physician trainees or a generalist practitioner, the intuitiveness of how to operate the AI is important. Abbreviations

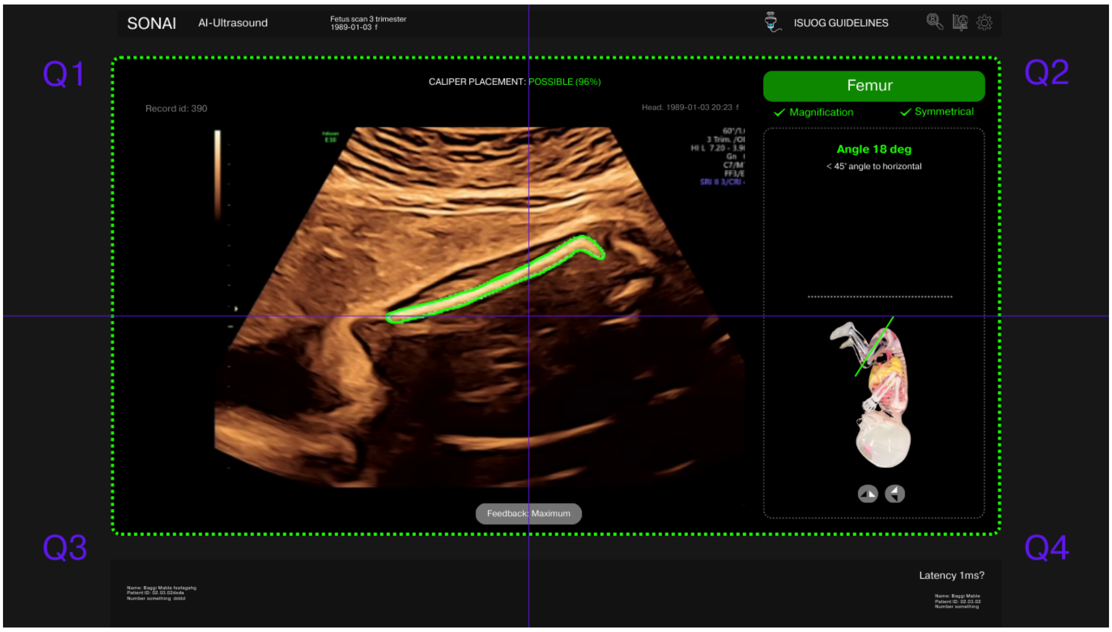


Fig. 3 A mock-up of our re-conceptualized version of the UI. For illustrative purposes, it has been divided into four quadrants. The view is of the Femur Standard Plane in a maximum feedback version.

and acronyms for instance contribute to more efficient flow of critical information in highly specialized medical departments or on the operating table, but have to align with the clinical practice in question. For an unspecialized physician or an unsupervised learner these can be more of a hindrance, and in interprofessional communication abbreviations can even pose a threat to patient safety (Tariq & Sharma, 2024). A more accessible and culturally neutral UI will not only make the AI more intuitive to use, but increase its overall capacity to "Ensure healthy lives and promote well-being (...)" (SDG3). We propose to display critically important UI elements with both non-abbreviated text, color and symbols/visuals to ensure compatibility and reduce the chance of misunderstandings.
For example the ultrasound display in Fig 3 is encompassed by a dotted green frame, the word "Femur" appears in the box that indicates which plane is at hand in Q2. Likewise the femur is outlined with green. To reduce eye travel distance, we propose to display all the UI elements of clinical relevance in Q2, including the checklist of a given standard plane, information on angle and image optimization etc. The selected UI design elements align with the ISUOG Practice Guidelines for ultrasound assessment of fetal biometry and growth (Salomon et al., 2019), the clinical standard of this device. We suggest that the guidelines can be read in the UI by toggling a screen element in the top of Q2. Ideally this would create a movable window pop-up containing the guidelines.
These UI elements would help the operator to confirm the criteria needed for objective evaluation of the biometric images, by illustrating each fulfilled criterion. In Fig 3 the operator can visually confirm if the femur is occupying more than half of the total image at an optimal angle. This is an ISUOG guideline that is materi-

alized into the AI via the "magnification" element in Q2 and in the checklist-box directly below it displaying the current angle with a color indication and text explanation. The AI feedback portrayed in the UI is thus contestable and rationally deductive by providing an explanation in situ of how it achieved its analysis of the ultrasound images.

5.2 Accountability

Friedman & Kahn (1992) argue that intentionality is a prerequisite for moral agency and that conceivable digital computer systems cannot achieve intentionality. From an ethics standpoint this means that only a human can hold the moral responsibility of consequences of an action performed by – or informed by the analysis of – a computer. Therefore, the AI should not withhold information on grounds of simplifying the UI (explicability), and it should not incontestably perform an actionable analysis (human autonomy).
In order to confidently trace the accountability of the outcomes of a decision augmented by the AI no further than the room it was made in, we see it as a necessary design requirement that all the information and reasoning that are possible to extract from the AI appears as UI elements. However, the participants and experts onboard the project have expressed a concern that a cluttered UI can be distracting and increase mental load, especially while a novice clinician gets a better grasp of the procedure. To accommodate these conflicting design requirements, we propose an option to toggle between minimum and maximum feedback, the latter being on by default. This would accommodate the accountability problem by making the minimum-feedback display an explicit choice. Fig 3 is an example of this proposal, and the toggleable UI-element for level of feedback is in the bottom of the frame at the intersection of Q3 and Q4.

5.3 Guidance system

When interviewing participants about what is on their wishlist for an EFW-AI-solution, a checklist of anatomical landmarks was mentioned. This could be a helpful feature for learners to confirm what is needed for each standard plane. This includes labels of the anatomical landmarks. Moreover, the participants asked for a sort of guidance system that can indicate how to locate the next standard plane, for example by recommending probe manipulation or changing settings on the ultrasound machine. In Q4 we suggest an avatar to help confirm if the operator has located a correct plane by indicating with a colored confirmation line. Based on a design suggestion from an experienced obstetrician, we propose adding buttons to flip the avatar around to change how the fetus is represented. This can increase the perception of how to locate the next standard plane for the operator, by mirroring how the fetus is presented in the womb.

6 Discussion

This study was designed to discover design requirements for the responsible design of a UI for an obstetric ultrasound AI. We employed a multifaceted qualitative ap-

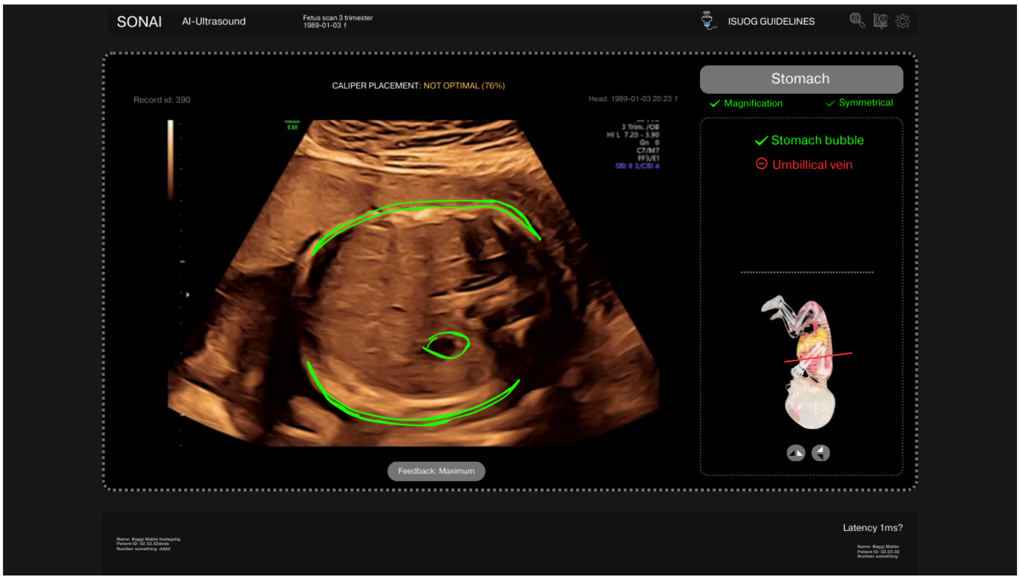


Fig. 4 Alternative example of AI-feedback showcasing a scan from the abdomen where the operator is not in the correct plane and where caliper placement isn't possible.

proach for empirical data collection exploring the needs and requirements of the potential users. By employing three different sources of values, we discovered implications of a relationship between the amount of information displayed on the screen and the perceived impact on clinical workflow and decision-making.
Fischer et al. (2023) performed a qualitative study of the perceptions of obstetrical clinicians on AI as a potential diagnostic tool for ultrasound. They find that an AI's ability to provide tools that strengthen a clinician's decision making rather than a strict focus on clinical outcomes should drive the design of AI solutions. This relates to the clinicians' trust in the AI's decisions, which can be achieved via traditional scientific validation (RCTs), professional endorsement, and explainability (Fischer et al., 2023). Our recommendations play into this by enhancing transparency, explicability, accountability, and user-friendliness through design adjustments of the maximum feedback layout.
It is difficult and sometimes counter-productive to formulate ethical principles directly into regulation and design requirements. For example, the four EU ethical principles are not expected to have an equal saying in every issue, and should be weighed against each other when there are conflicts (European Commission, 2019). In healthcare technology, the value of prevention of harm is likely to outweigh that of human autonomy e.g. via constant video surveillance of patients at high risk or the use of patient data to train machine learning models for better diagnosis. In law enforcement technologies however, human autonomy is likely to outweigh prevention of harm in the same technology examples due to the societal implications of constant video surveillance or crime prevention models trained on historical data (European Commission, 2019). As both scenarios are based on the respect of human values, the example shows that it is important to not only use a theoretical framework but to empirically investigate what weight to give a given

value over another. VSD principles hold significant promise for advancing the responsible and human-centered development of AI technologies in obstetrics and beyond, ultimately contributing to improved outcomes and experiences for both healthcare providers and patients.
According to Friedman & Kahn (1992) and Searle (1980), AI cannot be held responsible for actions. The responsibility is diffused and a symbolic head will be institutionally placed to be the target of the consequences. Integrating values into AI-UI design addresses moral accountability for clinicians and patients. Reverberi et al. (2022) found that over-reliance, under-reliance, or opaque AI judgments were not issues in their study. The doctors were able to build a correct mental model of the AI's error boundaries by interpreting AI output cues, similar to how they would evaluate a colleague's confidence. The key was providing intuitive and valid insights into the AI's reliability, which we propose to heighten in the AI of our study by adding additional certainty markers to the maximum feedback layout.
Our hybrid approach sources values relevant to the technology and context, ensuring the technology fits user needs and broader societal interests (Umbrello & van de Poel, 2021). These requirements can guide conceptual design frameworks. We must however acknowledge some limitations of our study design, within our method of sampling participants. Future research should aim to acquire insights into a broader participant base and include patients for a diverse clinical setting. Values and technologies are constantly changing. We need further user testing of the UI-design to fully understand how the AI-mediates values in the clinic as an ongoing process – the VtM model also encourages multiple cycles of anticipation and re-conceptualization (Smits et al., 2019, 2022). This study tested the two concepts of minimum and maximum feedback simultaneously, and although we articulated all of the discovered instances of negative value mediation with new design requirements, a return to the clinic with the next version of the prototype including patient investigations would be desirable.

7 Conclusion

In this study we examined the implications of value mediation by the UI of a new AI-based device supporting clinicians in performing fetal weight estimations. Our findings outlined in Table 1 and illustrated in Section 5 revealed how the positive and negative value experiences that come into play when testing the device in authentic settings can be translated into design requirements for the next iteration of the UI. The AI-supported procedure improved workflow efficiency by providing real-time feedback on standard plane optimization for accurate EFW assessments. The device contributed to clinical certainty as the confidence of the more novice physicians in identifying the correct standard planes grew. Conversely, our analysis revealed several negative value experiences that warrant attention for design optimization. For instance, concerns were raised about clinicians over-relying on AI feedback, gamifying scans to achieve maximum "green" indicators without ensuring accurate diagnostic information. This behavior compromises the values of human autonomy and explicability, as it shifts the workflow toward AI-dictation rather than AI-augmented decision-making. The central issue is that of opacity of AI reasoning, which the UI needs to display in a contestable manner in order to keep the moral accountability of the decisions made from the AI-augmented procedure with the clinician performing it. To address these findings, we propose a list of design recommendations focused on enhancing universability, usability, accountability and transparency.

1. *Number and content of UI-elements*: The AI-analysis ought to be intuitive and in real-time. Implement toggleable feedback modes (maxmimum → minumum) to accommodate different user preferences and skill levels.
2. *Transparency and accountability*: This can be achieved by explicitly referencing medical literature that the AI's reasoning and decision-making process are based upon – in this case the ISUOG guidelines. This strengthens both trust and accountability but also contestability. An example of implementation of this could simply be a toggleable info-box in the UI.
3. Guidance and learning support: A guidance system that helps the clinician develop a mental image of the positioning of the fetus and the standard planes in relation to each other is a highly requested feature for learners.
4. *Universal accessibility*: In addition to all the above, we recommend as a design standard that a color, text or a symbol never stands as the sole indicator of any UI element with diagnostic relevance.

These recommendations articulate the negative value mediation's we observed with the current prototype and incorporate AI-specific societal values into the design. They align with an overarching goal of promoting patient well-being and enhancing the effectiveness of healthcare delivery through responsible AI integration.

Acknowledgements This work was supported by the collaborative partners at The Pioneer Centre for AI and CAMES the Copenhagen Academy for Medical Simulation.

Author contributions statement A.S.M.P. and A.B.E conducted interviews, led the data analysis, and drafted the article. M.L.N guided the study and contributed with clinical insights and interpretation. All authors contributed to and reviewed the article..

Data availability Data supporting the findings of this study are available from the corresponding author, A.S.M.P. and A.B.E upon reasonable request. Complete transcripts of the interviews, beyond the supporting quotes, are however not publicly accessible due to confidentiality agreements with participants.

Ethics This study was reported for review to the Regional Ethics Committee of the Capital Region of Denmark and was deemed exempt from review, Case no: F-23073963. All participants were briefed on the details of the study, Anonymously, verbally agreed to be interviewed, and provided their informed consent to participate in the study by signature. The participant was offered the opportunity to opt out of the project.

The authors declare no competing interests.

References

Andreasen, L., Tabor, A., Nørgaard, L., Rode, L., Gerds, T., & Tolsgaard, M. (2021). Detection of growth-restricted fetuses during pregnancy is associated with fewer intrauterine deaths but increased adverse childhood outcomes: an observational study. *BJOG: An International Journal of Obstetrics & Gynaecology*, *128*(1), 77–85. Retrieved 2024-02-29, from https://onlinelibrary.wiley.com/doi/abs/10.1111/1471-0528.16380 (_eprint: https://obgyn.onlinelibrary.wiley.com/doi/pdf/10.1111/1471-0528.16380) doi: 10.1111/1471-0528.16380

Andreasen, L. A., Tabor, A., Nørgaard, L. N., Taksøe-Vester, C. A., Krebs, L., Jørgensen, F. S., ... Tolsgaard, M. G. (2021). Why we succeed and fail in detecting fetal growth restriction: A population-based study. *Acta Obstetricia et Gynecologica Scandinavica*, *100*(5), 893–899. Retrieved 2024-02-29, from https://onlinelibrary.wiley.com/doi/abs/10.1111/aogs.14048 (_eprint: https://obgyn.onlinelibrary.wiley.com/doi/pdf/10.1111/aogs.14048) doi: 10.1111/aogs.14048

Bohr, A., & Memarzadeh, K. (2020). The rise of artificial intelligence in healthcare applications. *Artificial Intelligence in Healthcare*, 25–60. Retrieved 2024-05-15, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7325854/ doi: 10.1016/B978-0-12-818438-7.00002-2

Charter of Fundamental Rights of the European Union. (2012, October). Retrieved 2024-06-03, from http://data.europa.eu/eli/treaty/char_2012/oj/eng (Legislative Body: EUMS) from https://www.nature.com/articles/s41599-021-00750-9 (Publisher: Palgrave) doi: 10.1057/s41599-021-00750-9

Dégallier-Rochat, S., Kurpicz-Briki, M., Endrissat, N., & Yatsenko, O. (2022, October). Human augmentation, not replacement: A research agenda for AI and robotics in the industry. *Frontiers in Robotics and AI*, *9*, 997386. Retrieved 2024-02-29, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9578547/ doi: 10.3389/frobt.2022.997386

Ethics guidelines for trustworthy AI | Shaping Europe's digital future. (2019, April). Retrieved 2024-05-20, from https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Fiorentino, M. C., Villani, F. P., Di Cosmo, M., Frontoni, E., & Moccia, S. (2023, January). A Review on Deep-Learning Algorithms for Fetal Ultrasound-Image Analysis. *Medical Image Analysis*, *83*, 102629. Retrieved 2024-05-15, from http://arxiv.org/abs/2201.12260 (arXiv:2201.12260 [cs, eess]) doi: 10.1016/j.media.2022.102629

Fischer, A., Rietveld, A., Teunissen, P., Hoogendoorn, M., & Bakker, P. (2023, October). What is the future of artificial intelligence in obstetrics? A qualitative study among healthcare professionals. *BMJ Open*, *13*(10), e076017. Retrieved 2024-06-06, from https://bmjopen.bmj.com/content/13/10/e076017 (Publisher: British Medical Journal Publishing Group Section: Obstetrics and gynaecology) doi: 10.1136/bmjopen-2023-076017

Friedman, B., & Hendry, D. (2019). *Value Sensitive Design: Shaping Technology with Moral Imagination.* doi: 10.7551/mitpress/7585.001.0001

Friedman, B., Kahn, P., & Borning, A. (2003, January). Value Sensitive Design: Theory and Methods.

Friedman, B., & Kahn, P. H. (1992, January). Human agency and responsible computing: Implications for computer system design. *Journal of Systems and Software*, *17*(1), 7–14. Retrieved 2024-03-18, from https://linkinghub.elsevier.com/retrieve/pii/016412129290075U (Number: 1) doi: 10.1016/0164-1212(92)90075-U

Goal 3 | Department of Economic and Social Affairs. (2012). Retrieved 2024-06-02, from https://sdgs.un.org/goals/goal3

Govaerts, M. J. B., Schuwirth, L. W. T., Van der Vleuten, C. P. M., & Muijtjens, A. M. M. (2011, May). Workplace-based assessment: effects of rater expertise. *Advances in Health Sciences Education*, *16*(2), 151–165. Retrieved 2024-02-29, from https://doi.org/10.1007/s10459-010-9250-7 doi: 10.1007/s10459-010-9250-7

Ihde, D. (1990). *Technology and the Lifeworld: From Garden to Earth.* Indiana University Press. Retrieved 2024-03-30, from https://philarchive.org/rec/IHDTAT-3

Kudina, O., & Verbeek, P.-P. (2019, March). Ethics from Within: Google Glass, the Collingridge Dilemma, and the Mediated Value of Privacy. *Science, Technology, & Human Values*, *44*(2), 291–314. Retrieved 2024-03-24, from https://doi.org/10.1177/0162243918793711 (Number: 2 Publisher: SAGE Publications Inc) doi: 10.1177/0162243918793711

Lieberman, H. (2009). User Interface Goals, AI Opportunities. *AI Magazine*, *30*(4), 16–22. Retrieved 2024-05-15, from https://www.proquest.com/docview/208118941/abstract/8877C8C9F54B4280PQ/1 (Num Pages: 7 Place: La Canada, United States Publisher: Association for the Advancement of Artificial Intelligence Section: Articles) doi: 10.1609/aimag.v30i4.2266

Meske, C., & Bunde, E. (2023, April). Design Principles for User Interfaces in AI-Based Decision Support Systems: The Case of Explainable Hate Speech Detection. *Information Systems Frontiers*, *25*(2), 743–773. Retrieved 2024-05-15, from https://doi.org/10.1007/s10796-021-10234-5 doi: 10.1007/s10796-021-10234-5

Mintz, Y., & Brodie, R. (2019, March). Introduction to artificial intelligence in medicine. *Minimally Invasive Therapy & Allied Technologies*, *28*(2), 73–81. Retrieved 2024-05-15, from https://www.tandfonline.com/doi/full/10.1080/13645706.2019.1575882 doi: 10.1080/13645706.2019.1575882

Nicholls, D., Sweet, L., & Hyett, J. (2014). Psychomotor Skills in Medical Ultrasound Imaging. *Journal of Ultrasound in Medicine*, *33*(8), 1349–1352. Retrieved 2024-02-29, from https://onlinelibrary.wiley.com/doi/abs/10.7863/ultra.33.8.1349 (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.7863/ultra.33.8.1349) doi: 10.7863/ultra.33.8.1349

Paladini, D. (2009). Sonography in obese and overweight pregnant women: clinical, medicolegal and technical issues. *Ultrasound in Obstetrics & Gynecology*, *33*(6), 720–729. Retrieved 2024-02-29, from https://onlinelibrary.wiley.com/doi/abs/10.1002/uog.6393 (_eprint: https://obgyn.onlinelibrary.wiley.com/doi/pdf/10.1002/uog.6393) doi: 10.1002/uog.6393

Reverberi, C., Rigon, T., Solari, A., Hassan, C., Cherubini, P., & Cherubini, A. (2022, September). Experimental evidence of effective human–AI collaboration in medical decision-making. *Scientific Reports*, *12*(1), 14952. Retrieved 2024-06-06, from https://www.nature.com/articles/s41598-022-18751-2 (Publisher: Nature Publishing Group) doi: 10.1038/s41598-022-18751-2

Salomon, L. J., Alfirevic, Z., Da Silva Costa, F., Deter, R. L., Figueras, F., Ghi, T., ... Yeo, G. (2019, June). ISUOG Practice Guidelines: ultrasound assessment of fetal biometry and growth. *Ultrasound in Obstetrics & Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology*, *53*(6), 715–723. doi: 10.1002/uog.20272

Searle, J. R. (1980, September). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–424. Retrieved 2024-06-06, from https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/abs/minds-brains-and-programs/DC644B47A4299C637C89772FACC2706A doi: 10.1017/S0140525X00005756

Smits, M., Bredie, B., Goor, H. v., & Verbeek, P.-P. (2019, November). Values that Matter: Mediation theory and Design for Values. In *Research Perspectives in the era of transformations: Conference proceedings* (pp. 396–407). Academy for Design Innovation Management. Retrieved 2024-03-18, from https://research.utwente.nl/en/publications/values-that-matter-mediation-theory-and-design-for-values doi: 10.33114/adim.2019.03_203

Smits, M., Ludden, G., Peters, R., Bredie, S. J. H., van Goor, H., & Verbeek, P.-P. (2022). Values that Matter: A New Method to Design and Assess Moral Mediation of Technology. *Design Issues*, *38*(1), 39–54. Retrieved 2024-02-29, from https://search.ebscohost.com/login.aspx?direct=true&db=buh&AN=154440862&site=ehost-live (Publisher: MIT

Press) doi: 10.1162/desi_a_00669

Tariq, R. A., & Sharma, S. (2024). Inappropriate Medical Abbreviations. In *StatPearls*. Treasure Island (FL): StatPearls Publishing. Retrieved 2024-04-07, from http://www.ncbi.nlm.nih.gov/books/NBK519006/

Tolsgaard, M. G., Boscardin, C. K., Park, Y. S., Cuddy, M. M., & Sebok-Syer, S. S. (2020, December). The role of data science and machine learning in Health Professions Education: practical applications, theoretical contributions, and epistemic beliefs. *Advances in Health Sciences Education*, *25*(5), 1057–1086. Retrieved 2024-02-29, from https://doi.org/10.1007/s10459-020-10009-8 doi: 10.1007/s10459-020-10009-8

Topol, E. J. (2019, January). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, *25*(1), 44–56. Retrieved 2024-02-29, from https://www.nature.com/articles/s41591-018-0300-7 (Number: 1 Publisher: Nature Publishing Group) doi: 10.1038/s41591-018-0300-7

Umbrello, S., & van de Poel, I. (2021). Mapping Value Sensitive Design onto AI for Social Good Principles. *AI and Ethics*, *1*(3), 283–296. (Number: 3)

van de Poel, I. (2013). Translating Values into Design Requirements. In D. P. Michelfelder, N. McCarthy, & D. E. Goldberg (Eds.), *Philosophy and Engineering: Reflections on Practice, Principles and Process* (pp. 253–266). Dordrecht: Springer Netherlands. Retrieved 2024-03-30, from https://doi.org/10.1007/978-94-007-7762-0_20 doi: 10.1007/978-94-007-7762-0_20

Yi, J., Kang, H. K., Kwon, J.-H., Kim, K.-S., Park, M. H., Seong, Y. K., ... Bang, W.-C. (2021, January). Technology trends and applications of deep learning in ultrasonography: image quality enhancement, diagnostic support, and improving workflow efficiency. *Ultrasonography (Seoul, Korea)*, *40*(1), 7–22. doi: 10.14366/usg.20102