

Tallinn University
Cyprus University of Technology

School of Digital Technologies
Department of Multimedia and Graphic Arts

Keeping the Human in the Loop

A Case Study in the Medical Domain, that Should Increase Trust in
Workflows Supported by Artificial Intelligence.

Master Thesis

Svenja Lisa Dittrich

Author: " "
Supervisor: " "
Supervisor: " "
Head of the School: " "

Author's declaration and non-exclusive licence for the use of the thesis


I, Svenja Lisa Dittrich

1. Have compiled the Master's Thesis named in Clause 2 independently. The research of other authors, important opinions from literature and other sources are cited.
2. Grant Cyprus University of Technology and Tallinn University a permit (a non-exclusive licence) without claiming payment of remuneration to:
 - 2.1. Reproduce for the purpose of conservation and electronic publication in the repository of Cyprus University of Technology Academic Library;
 - 2.2. Make public in the repository of Cyprus University of Technology Academic Library the following thesis created by me:

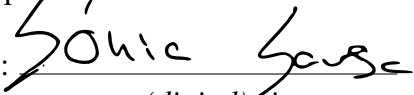
Keeping the Human in the Loop - A Case Study in the Medical Domain, that Should Increase Trust in Workflows Supported by Artificial Intelligence.


supervised by Sonia Sousa and Dr. Alexander Wiethoff.

3. The permissions granted in Clause 2 shall be given from the evaluation of the thesis with positive result until the term of the protection of copyright.
4. I am aware of the fact that the author also retains the rights mentioned in Clause 2.
5. I certify that granting the non-exclusive licence does not infringe the intellectual property rights of other persons or the rights arising from the legal acts regulating the protection of natural persons upon processing of personal data.

The author of the thesis: _____  07.05.2021
(digital) signature and date

The student is allowed to proceed to the defence of the final thesis.

Supervisor (Sonia Sousa):  _____
(digital) signature and date

Supervisor (Dr. Alexander Wiethoff):  _____ 07.05.2021
(digital) signature and date

Abstract

The technical advancement of medical image processing, in the last decades, can be attributed to improvements in imaging technologies. These technical advancements have meant both more images per individual scan and a greater demand for scans as a diagnostic tool. Thus radiologists today face an ever-increasing amount of medical images to review. As machine learning enhances its applicability and significance in the medical-imaging domain, AI applications still lack acceptance by radiologists due to insufficient integration into the clinical workflows. Therefore we designed a case study that aims to provide transparency related to the interpretation and control of model output to minimize disruptions to professional expertise and work cultures. The goal of the study is to measure the effectiveness of a design solution for a new machine learning process. With the new development we predict the increase of efficiency through transparency in the clinical workflow. This will have a positive impact on radiologists' trust and acceptance in Artificial Intelligence.

The evaluation results conclude that the new design lets the user achieve their goals quicker and easier. This assertion is strengthened by an increase in the perception of usability and task success as well as a decrease of time on task. Measuring trust revealed an improvement related to benevolence and the assessment of competence in the new design. The outcome provides essential input for future developments, not only in the implementation of medical solutions but also supply valuable insights for the Human Computer Interaction (HCI) community. The authors provide design recommendations regarding the design process as well as a checklist including AI-Cues. Following the suggested framework should help create transparent connections that keep the human in the loop working with AI-supported user flows. These loops should prevent the technology from being related to a black box that users can't trust.

Keywords: User-centered, Workflow, Machine-learning, Artificial Intelligence, Structured Reporting, Medical Imaging, Trust, Acceptance

Acknowledgements

I would first like to especially thank my thesis advisor Sonia Sousa of the Tallinn University and Dr. Alexander Wiethoff, at the Ludwig Maximilians University of Munich. Both guided me through every step of this thesis providing valuable feedback, support, and far most inspiration. I especially want to thank Dr. Alexander Wiethoff, my mentor through all my studies and at work, who introduced me to deepc and guided me in the direction of this thesis.

I sincerely like to thank the team of deepc, supporting me with this thesis. Especially Julia Moosbauer, Franz Pfister, Tristan Post, and Paul Mayer, providing valuable domain knowledge and helping me in the recruitment phases. Furthermore, I would like to thank the experts involved in the interviews and the validation process for this research project: Sarah Schläger, Ulrike Grimm, Bene Wiestler, and Felix Hofmann. Without their participation and input, the research could not have been successfully conducted.

Finally, I must express my very profound gratitude to my partner, Fabian Nußberger, for providing me with unfailing support, never-ending patience, and continuous encouragement throughout my years of study, especially through researching and writing this thesis. Also, many thanks to my little Emil, who patiently let me finish writing my thesis.

This accomplishment would never have been possible without all of them. Thank you so much for your support.

Svenja Dittrich

Contents

1	Introduction	12
1.1	Research Problem and Significance	13
1.2	Research Goal and Motivation	13
1.3	Research Question	15
1.3.1	Main Research Question	15
1.3.2	Sub Research Questions	15
1.4	Theories and Methods	15
1.5	Plan of Implementation	19
1.6	Expected Outcomes	20
2	Literature Review	21
2.1	Artificial Intelligence	21
2.1.1	Levels of AI	22
2.1.2	AI in Medical Image Interpretation	22
2.2	Radiologists Workflow	23
2.2.1	Image Interpretation	24
2.2.2	Reporting	25

2.3	Trust in Human Computer Interaction	26
2.3.1	Cues	27
2.3.2	Human Computer Trust Model	28
2.4	Trust in AI Solutions in the Medical Domain	28
2.4.1	Opinions	29
2.4.2	Barriers	29
2.5	Visualization of AI and Machine Learning Processes in HCI	30
2.5.1	Cues and Actions	30
2.5.2	Transparent Design	30
2.5.3	Explainable Design	31
2.5.4	Human Computer Collaboration	32
3	Discover and Define	33
3.1	Participatory Research Design	33
3.2	Interviews	37
3.3	Data Analysis	38
3.3.1	Observation	38
3.3.2	Survey	40
3.3.3	Interviews	41
3.3.4	Summary and Resulting Personas	44
4	Design Development	46
4.1	Design Adaptions	47
4.1.1	Usability and Workflow Adaptions	48

4.1.2	Design for Trust	53
4.1.3	Design for Transparency	59
4.2	Human in the Loop	62
4.3	Formative Evaluation	63
5	Summative Study	64
5.1	Instruments and Methods	65
5.2	Participants	67
5.3	Study	67
5.3.1	Risks to Validity	67
5.3.2	Pilot Test	68
5.4	Findings	68
5.4.1	Task Success and Time on Task	68
5.4.2	Survey	71
5.4.3	Qualitative Questions	73
5.5	Design Recommendations	75
6	Discussion	78
7	Conclusion	80
8	References	81
9	Appendix	86
9.1	User Study Plot: Expert Interviews	86

9.2 User Study Plot: Summative Study V A+B 102

9.3 Call for Participation (Summative Study) 116

Chapter 1

Introduction

Due to the rapid technical progress over the last decades, the complexity of imaging procedures in the medical domains is rising. Through new technologies and increasing numbers of diagnostic images, radiologists today face an increased amount of medical images to process in their daily work routine. While the technology has changed, the tasks, that a radiologist has to perform stayed roughly the same. The increased amount of complexity, combined with shorter reading time due to a higher workload, comes at a high cost. A higher reading speed can lead to more diagnostic errors. Radiologists not only suffer from time pressure but also from fatigue and many different cognitive biases such as satisfaction of search or inattention blindness (Williams et al. (2020)). These phenomena can lead to false diagnoses, threaten its credibility and finally threaten the patients' health.

Diagnostic tools trained by machine learning can aid radiologists in their daily work routine (e.g. by finding anomalies or streamline reporting) and alleviate the aforementioned problems. Thus Machine learning processes gain usage and significance in the medical imaging domain. Therefore radiology workflows, with AI model integration, will become increasingly critical to support the process of interpreting medical images (Dikici, Bigelow, Prevedello, White, and Erdal (2020)). Due to the complexity of machine learning processes and the use of artificial intelligence, users are not always in the picture. It is hard to understand the abstract processes and interpret the outcome of AI solutions. Therefore AI is often presumed as a "black box" to the user which could also lead to uncertainties in the credibility of AI processes (Strohm, Hehakaya, Ranschaert, Boon, and Moors (2020)). This lack of transparency can lead to distrust and lower acceptability and adoption of AI-based diagnostic tools.

1.1 Research Problem and Significance

As machine learning enhances its applicability and significance in the medical-imaging domain, AI applications still lack acceptance by radiologists due to insufficient integration in clinical workflows (Dikici et al. (2020); Strohm et al. (2020); Yu and Kohane (2019)). Strohm et al. (2020) identified this lack of acceptance and trust, as one of four major barriers for integrating AI applications. It appears alongside the problem of missing strategies for integrating AI applications in the clinical workflow. This finding is being strengthened by Yu and Kohane (2019) pointing out that trust in medical technology is closely related to its anticipated utility. Disruptions in the workflow may cause this trust to break off. Yu and Kohane (2019) also points out that even robust AI applications can reduce efficiency and cause additional medical errors if not adequately integrated into the current clinical workflow. According to Dikici et al. (2020), this disruption can prevent the flow of high-quality data. One cause for this disruption can be the lack of transparency, involvement of radiologists, and their control in the machine learning process. (“Impact of artificial intelligence on radiology: a EuroAIM survey among members of the European Society of Radiology” (2019); *Med Students and AI June 2019 — American College of Radiology* (n.d.))

A robust clinical system requires multiple quality control steps which become a larger problem when dealing with more complex 4D or 5D data (*Med Students and AI June 2019 — American College of Radiology* (n.d.)). There has been less focus on real-world implementation and the associated challenges to transparency and accountability (Sendak et al. (2019)). Thus, the European Society of Radiology (ESR) demands that radiologists must play a leading role in developing and validating AI applications to medical imaging to enhance the quality of image processing. (“Impact of artificial intelligence on radiology: a EuroAIM survey among members of the European Society of Radiology” (2019))

1.2 Research Goal and Motivation

The goal of the study is to:

- Use trust values to leverage radiologists’ acceptance into a structured medical reporting workflow, that was supported by AI.

We aim to achieve this by using a participatory design approach that includes radiologists in the process early on. It also includes the design of an interface for a machine learning solution that can be easily integrated into a radiologists clinical workflow.

The design focuses on the task of contextual and structured image reporting. The prototype should provide transparency about AI and its generated output as well as allow for control of the model output. This way we aim to minimize disruptions to professional expertise and work cultures. Preliminary expert interviews provide qualitative insights into opportunities and barriers radiologists are facing in their workflow. A subsequent study provides data to validated the proposed solution.

1. **Objective I:** Design for transparency: Measure the effectiveness of the design for transparency (values - visual cues, social cues, content cues) to increase radiologists' trust in the diagnostic tool.
2. **Objective II:** Understand how trust interrelates with acceptance of the diagnostic tool. (trust vs transparency - workflow, visual cues, social cues, content cues)
3. **Objective III:** Provide additional insights on how to design towards a goal of fostering trust and acceptance in Artificial Intelligence.

To fulfill the objectives we designed, in collaboration with radiologists, an AI-driven structured reporting solution. We integrated ideas into the new design for increasing trust in the AI output. Furthermore, we measure if these changes lead to more acceptance of the AI generated findings by the radiologists. Finally we will use our findings to formulate recommendations on how to design for trust to leverage acceptance in AI.

1.3 Research Question

1.3.1 Main Research Question

Fulfilling the previously mentioned objectives should help to answer the following research questions:

How can machine learning processes be designed to foster trust in users?

1.3.2 Sub Research Questions

1. *How can we design a transparent workflow integrated reporting solution (using trust values like visual, content and social cues) for radiologists, supported by artificial intelligence?*
2. *What makes radiologists trust the output of a reporting artificial intelligence solution and increases their acceptance using it?*
3. *To what extent can trust be a key to leverage effective artificial intelligence structured reporting work-flow?*

1.4 Theories and Methods

A participatory user-centered design approach, based on the double diamond principle (see figure 1.1), provided important input for the development of the prototype. A solution, based on real users' needs, should be created in the end. In the "discover" phase, expert interviews with four participants from the medical domain (focusing on radiology) provided information on the status quo, current needs, visions, and barriers. A clickable prototype of an AI-driven operating system (deepcOS), developed by deepc was used as a first basis, combined with a survey related to the perception of trust and usability. Semi-structured interviews provided further contextual information. These resulting findings were analyzed in the "define" phase to create a baseline, define requirements and goals for the prototype development in the "develop" phase. Further, the new clickable prototype was validated in a summative evaluation later on in the "deliver" phase. Quantitative A/B testing was conducted with 10 users using an online evaluation tool called "Loop11"¹. The testing application provided an

¹<https://www.loop11.com>

evaluation process for clickable prototypes. To compare the new design to the original prototype, that was used in the expert interviews, the same surveys related to the perception of trust and usability were used. Additionally we measured task completion time and the task success rate. A subsequent structured interview provided additional context information which will provide important insights for future research and developments. This study design was chosen to provide a triangulation of data that should help to focus on a user-centered and participatory design approach. A detailed overview of steps, connected research questions, and methods can be viewed in the following table.

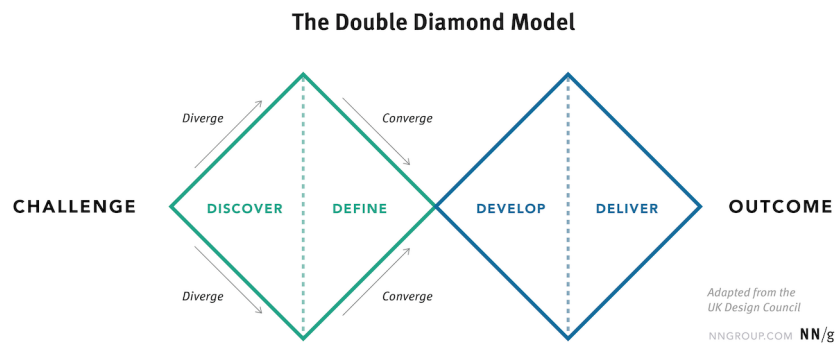


Figure 1.1: The Double Diamond Model by NNGroup²

Steps of Research	Research Question (RQ) / Sub-question (SQ) / Objectives (OB)	Method
<p>Step 1 Discover Theoretical background on AI, radiologist workflows and trust.</p>	<p>SQ1 How can we design a transparent workflow integrated reporting solution (using trust values like visual, content and social cues) for radiologists, supported by artificial intelligence?</p> <p>SQ3 To what extent can trust be a key to leverage effective artificial intelligence structured reporting work-flow?</p>	<p>Literature Review The literature review aims to understand the basics of artificial intelligence, workflows in the medical domain and the importance of trust in the field of human computer interaction.</p> <p>Expert Interviews This should provide a baseline to start the experiment and understand current needs, fears and expectations from real users.</p>
<p>Step 2 Define Synthesizing the information from the discovery phase into a problem definition and define the design idea and features.</p>	<p>SQ2 What makes radiologists trust the output of a reporting artificial intelligence solution and increases their acceptance using it?</p> <p>OB II Understand how trust interrelated with artefact acceptance.</p>	<p>Data Analysis using a code book to collect and compare the data and find out about risks, opportunities, features, needs and expectations. Resulting personas should provide defined goals from the user perspective.</p>
<p>Step 3 Develop A design proposal as a clickable prototype.</p>	<p>-</p>	<p>A Design Proposal, based on deepcOS and the outcomes of the expert interviews</p>

<p>Step 4 Deliver Results of the evaluation and discussion.</p>	<p>RQ How can machine learning processes be designed to foster trust in users?</p> <p>OB I Measure how effective is the design for transparency.</p> <p>OB III Provide additional insights on how to design to foster trust and acceptance in Artificial Intelligence.</p>	<p>The Summative Study and Analysis aims to investigate the impact of the developed experiment on users trust in the use of Artificial Intelligence.</p>
--	---	---

1.5 Plan of Implementation

For the implementation of our study, we considered several resources. The medical startup deepc provided the software basis for this research, a company that implemented deepcAI, an AI algorithm for interpreting CT images. Based on their software, deepcOS, we created a clickable prototype. We chose a set of tasks based on the process of documenting a finding (medical anomaly) in a brain CT scan.

Study plots with surveys and interview questions, used for the expert interviews, were prepared. For the development of the final prototype, the design software "Figma" was used. As mentioned before, due to the ongoing COVID19 pandemic, it was not possible to visit the medical staff in their natural working environment. To overcome this issue, we used the conferencing tool "Google Meet" to conduct the expert interviews. More detailed information about the process can be found in chapter "3 Discover and Define". The summative testing of the prototype was integrated into an online testing platform, "Loop11". Thus users could quickly evaluate the prototype in a remote and unmoderated session without time or location restrictions. Participants for both phases needed to be recruited. We invited four individuals who work in the radiology department for expert interviews. The target number of users for the summative testing phase was ten individuals, preferably from the medical domain. The leading target user group were radiologist, but due to limited availability, the participation was extended to potential users in the health system or healthcare industry. An overview of the time span can be viewed under figure 1.2

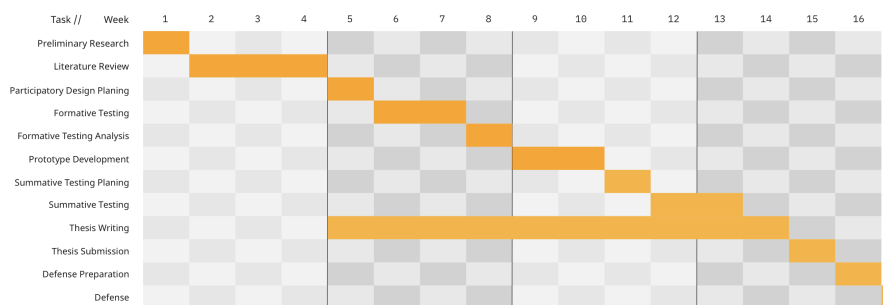


Figure 1.2: Gantt Chart for planning the implementation of the thesis

1.6 Expected Outcomes

The study's goal is to measure the effectiveness of a design solution for a new machine learning process. We predict the increase of efficiency through transparency. The new design will positively impact radiologists' trust and acceptance in the use of Artificial Intelligence solutions. The outcome should provide input for future developments in implementing medical solutions and provide as well some insights for the HCI community, how to design trustful, user-friendly, transparent solutions supported by artificial intelligence.

Chapter 2

Literature Review

The literature review included a scoping study methodology (Levac, Colquhoun, and O'Brien (2010)). Papers and proceedings were searched on Springer, in the ACM Library and Google Scholar containing "Artificial Intelligence" OR "machine learning" in the medical field, furthermore papers were selected based on additional tags like "trust", "workflow radiologists" and selected based on titles of papers. Additionally a backwards search provided additional papers and proceedings. The references of promising papers were scanned and selected this way.

Due to the fast development of the domain and community, only papers of the last years (five years max) were chosen for the literature review to work with the latest findings. The next paragraphs provide a summary of the outcomes.

2.1 Artificial Intelligence

The first chapter of this literature review should provide some contextual input about Artificial Intelligence (AI). It helps to create a baseline for further research and discussions. Due to the extensive literature offered by the community of computer science, the author narrowed it down to context-related topics.

2.1.1 Levels of AI

Chartrand et al. (2017) describe AI as a branch of computer science devoted to creating systems that perform tasks that ordinarily require human intelligence. As a subfield of AI, machine learning processes are used to train computers to perform tasks without explicit programming, e.g., distinguishing data patterns. Furthermore, representation learning can be seen as a type of machine learning where the algorithm learns the best features to classify the provided raw data. The deepest level of AI that should be considered here is the level of deep learning. These deep learning systems propose an end-to-end approach by learning simple features as components of more complex characteristics, such as shapes, lesions, organs, therefore leveraging the compositional nature of images (Chartrand et al. (2017)). In general, the methods to train these models can be divided into supervised and unsupervised learning algorithms. In supervised learning, a model is presented with a data set of input features and labels. The target is then to interpret and predict the input source based on the given model parameters (Litjens et al. (2017)). Unsupervised learning algorithms process data without labels and are trained to find patterns, such as latent subspaces. For our further research, we focus on machine learning processes that follow a supervised structure. Users need to feed the AI, first with different input modalities, to train it until it continues autonomously by learning unsupervised.

2.1.2 AI in Medical Image Interpretation

Over the past few years, the emerging fields of AI and deep learning have attracted much attention in the medical domain, especially in the community of radiologists (*5 key takeaways from a patient survey about AI in radiology* (n.d.)). The most successful AI models are used for image analysis, especially convolutional neural networks (CNN). CNNs contain many layers that transform their input with convolution filters of a small extent. Work on CNNs has been done since the late seventies (Fukushima and Miyake (1982)), and they were already applied to medical image analysis in 1995 by Lo, Lou, Chien, and Mun (1995). Today they are the standard technique in exam classification and the preferred method for object or lesion classification. In exam classification, one typically has one or multiple images as input with a single diagnostic variable as output (e.g., disease present or not) (Litjens et al. (2017)). Further areas of successful image interpretation are observed in the detection of organs, region and landmark localization, object or lesion detection, segmentation, lesion segmentation, and registration. In contrast to classification and segmentation, the research community seems

not yet settled on the best way to integrate deep learning techniques in registration methods. Not many papers have appeared yet on this topic, and existing ones have distinctly different approaches (Litjens et al. (2017)). In various studies, it has been shown that to train more complex models of AI, it is significant how many experts annotated data were used (Dikici et al. (2020)) to evaluate the quality of image interpretation when using AI models.

2.2 Radiologists Workflow

As pointed out by Dikici et al. (2020), the process of implementing and maintaining the standard radiology workflow can be simplified described in five main steps (2.1).

The images acquired in digital imaging and communications in medicine (DICOM) format are sent to a DICOM router, a configurable framework capable of sending/receiving DICOM images to and from a predefined address. The router sends the images to the picture archiving and communication system (PACS) and vendor-neutral archive (VNA), which is a technology enabling the storage of medical images in a standard format and offering a generic interface, thereby making the data accessible to multiply healthcare professionals regardless of the type of system from which images are originating. Using dedicated workstations, radiologists access image data stored in PACS for study visualizing, post-processing and interpretation. Nonradiology clinicians may also view medical images stored only in a VNA through links in each patient's electronic medical record (EMR).

1. Medical images are acquired with a standards morality (e.g. CT, MRI, etc.) by a technologist.
2. The images acquired in digital imaging and communications in medicine (DICOM) format are sent to a DICOM router, a configurable framework that is capable of sending/receiving DICOM images to and from a predefined address.
3. The router sends the images to the picture archiving and communication system (PACS) and vendor-neutral archive (VNA) which is a technology enabling the storage of medical images in a standard format and offering a generic interface, thereby making the data accessible to multiply healthcare professionals regardless of the type of system from which images are originating.
4. Using dedicated workstations, radiologists access image data stored in PACS for study visualization, post-processing, and interpretation.

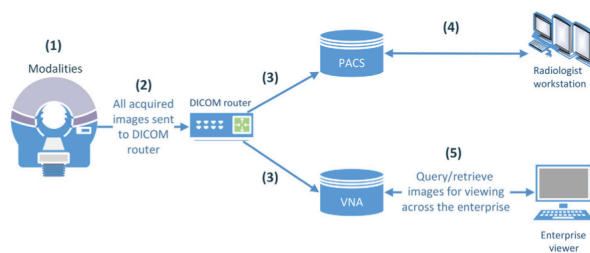


Figure 2.1: A simplified view of an exemplary radiology workflow by Dikici et al. (2020)

5. Nonradiology clinicians may also view medical images stored only in a VNA through links in each patient’s electronic medical record (EMR).

2.2.1 Image Interpretation

Despite the introduction of digital PACS globally over the past few decades and the existence and acceptance of international "Digital Imaging and Communications in Medicine" (DICOM) standards for the storage and transfer of medical imaging data, there remain significant barriers to large-scale big data sharing. Not all clinics have an infrastructure that allows interoperability with other systems. Additionally, DICOM standards often come with substantial variations in the quality of meta tags and other data points (Harvey and Glocker (2019)). The imaging studies are not standardized and may differ for each vendor’s hardware at the same clinical site. Gueld et al. (2002) found that it was impossible to automatically categorize medical images based solely on their DICOM meta tags as around 15 % of all studies were mislabeled due to human factors (Harvey and Glocker (2019)). In a paper on data readiness, Neil Lawrence proposed a three-point scale to allow inter-disciplinary conversations on inherent readiness of data. Inspired by this scale, Harvey and Glocker (2019) propose a four-point medical imaging data readiness (MIDaR scale) which takes into account additional and specific requirements related to research on medical imaging data.

1. **Level D:** Representing only its initial intended purpose of acting as a record of clinical activity with no further consideration of research of any kind (e.g. patient identifiable information).
2. **Level C** Anonymised and accessible via ethical approval, data extraction and access control.
3. **Level B:** The quantity of relevant data sets are fully accounted for, and large scale errors in data structure and format have been resolved (outcome of data selection from Level C)

4. **Level A:** Structured, fully annotated with minimum noise and contextually appropriate and ready for a specific ml task.

It is a long way towards level A data where its value increases but the volume decreases. Various data labeling processes must be undertaken to create strong labels. Standard techniques include natural language processing (NLP) for information extraction, expert radiologist manual contouring, derivation of consensus opinions, or lineage to existing external clinical gold standard results (Harvey and Glocker (2019)).

2.2.2 Reporting

Traditional narrative spoke reports are associated with excessive variability in the language, length, and style, which can minimize report clarity and make it difficult for referring clinicians to identify the critical information needed for patient care. The DICOM Structured Reporting specification has been in use for over 20 years (Bidgood (1998)), describing the technical requirements of integration data elements in radiology reports. Different definitions can be found, despite its broad application. First, it can refer to standardized reporting where a template with required components is used to create a structured report which improves the quality by ensuring coherent and complete reports. The second definition refers to how the content is arranged, integrated into the workflow, improving usability and acceptance (Olthof, Leusveld, de Groot, Callenbach, and van Ooijen (2020)).

A drawback of structured reporting is the inflexibility of reporting where radiologists are limited to a given structure and might not apply to their situation/case. In further developments, structured contextual reporting has proved itself beneficial, where different templates can be used and combined. Olthof et al. (2020) developed a template for structured contextual reporting, tailored to the clinical scenario, where the radiologists can use modular building blocks to customize a report to particular situations for context-dependent structured reporting for both content and structure. The template can be activated through voice command, anywhere in a radiology report. Olthof et al. (2020) base themselves on a technical paper on Management of Radiology Report Templates (MRRT) ? point out the application of templates in the exchange of information in clinical practice and support their notions that it may improve efficiency and quality in radiology reporting processes.

2.3 Trust in Human Computer Interaction

Trust has been, over the years, a topic of interest across many different domains such as philosophy, psychology, economics, sociology, management, marketing, and others, who have all defined and studied the concept in their way. Due to the broad field of research on trust, we focus on the definition provided by Rousseau, Sitkin, Burt, and Camerer (1998). Define it as a “psychological state comprising the intention to accept vulnerabilities based upon positive expectations of the intentions or behavior of another” (p.395). Their work was influenced by Mayer, Davis, and Schoorman (1995) who defines trust as “willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party” (p.712).

Trust is also essential when it comes to the field of human-computer interaction (HCI). Especially when we look beyond traditional HCI constructs like usability, trust represents a vital factor when interacting with technologies or, in our case, with artificial intelligence. Due to Morville (n.d.) credibility is one of eight major essential components in UX design (see figure 2.2). Depending on the product’s goals and user needs, a design can focus and follow different purposes. In our case, we focus on credibility since we want to find out how trust can be enhanced in connection with AI-supported workflows.

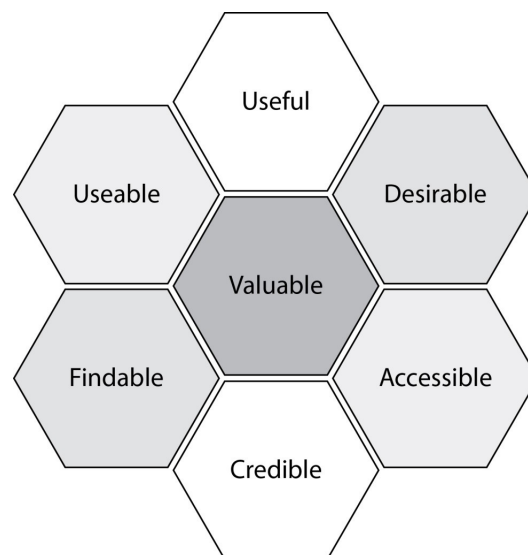


Figure 2.2: Honeycomb of UX Design by Morville (n.d.) ¹

2.3.1 Cues

The prevailing cognitive model of trust assumes that users search for information on risks and benefits and weigh them against each other to reach a decision. Specific cues need to be considered for this assessment. Wang (2009) allocates most trust components in three major dimensions, visual, content, and social-cue design. The focus of the research community on this topic is highly related to online and web environments. We assume that parts of these findings can be adapted to the general domain of human-computer interaction.

Visual Design

Due to the framework of trust-inducing design features by Wang (2009) and Fimberg and Sousa (2020), visual design defines the graphical design aspect and the structural organization of displayed information. It aims to shape and improve the user experience by considering the effects of different design elements on the usability of products and their aesthetic appeal. Some examples of design features can be: implementation of easy-to-use navigation (simplicity and consistency) or navigation reinforcements (e.g., guides, tutorials, instructions).

Content Design

Content design refers to the informational characteristics of components that are integrated into interfaces. These could be of textual or graphical nature, whereby the visual facet is highly connected to the domain of visual cue design. One should use these design features to communicate comprehensive, correct, and current product information.

Social-cue Design

Social-cue Design relates to embedding social and interpersonal cues, such as social presence and face-to-face interaction, into different communication media. Social cues can be visualized by including representative photographs or video clips or the use of synchronous communication media (e.g., instant messaging, chat lines, video telephony).

2.3.2 Human Computer Trust Model

Keeping in mind that trust affects so many facets of user interaction with technology, measuring it is a valuable endeavor. Therefore Gulati, Sousa, and Lamas (2019) proposed a Human-Computer Trust Model (HCTM) consisting of a set of twelve attributes (namely motivation, willingness, competence, benevolence, predictability, honesty, and reciprocity) to ascertain the essence of trust and help to measure it in user technology. Their latest study results concluded that three attributes, namely benevolence, competence, and perceived risk, are statistically significant (Gulati et al. (2019)).

Benevolence is understood as the technology being able to provide adequate, effective, and responsive help to the end-user, which eventually helps the user attain specific goals related to their interaction (Gulati et al. (2019), Wang (2009)). Individuals who perceive that technology can provide the needed help will perceive fewer risks and uncertainties associated with its use, which can eventually translate into higher trust and continual use of the technology (Mcknight, Carter, Thatcher, and Clay (2011)).

Competence of a system that is a direct representation of whether or not it has all the features and the functionalities to perform its intended tasks (Mcknight et al. (2011)). Essentially, if a user perceives technology to be competent (i.e., having all desired functionality to achieve a particular outcome), then there is a high likelihood that they would place trust and act on the advice and recommendations offered by the technology (Gulati, Sousa, and Lamas (2018)).

Gulati et al. (2018) perceive risk as a subjective assessment on the part of the end-user of the probability of a specified type of incident happening when using a technical artifact and how concerned they are with the consequences of their action. As earlier described, the notion of perceived risk particularly becomes essential when interacting with AI systems with a black-box nature.

The final version of the HCTS (Human Computer Trust Scale), by Gulati et al. (2019), can be viewed in the appendix.

2.4 Trust in AI Solutions in the Medical Domain

As machine learning enhances its applicability, significance, and potential in the medical-imaging domain, AI applications still lack acceptance by radiologists due to insufficient integration in the clinical workflows and other barriers (Strohm et al. (2020); Yu and Kohane (2019)).

Both even identified the significant variance in acceptance and trust as one of four major barriers to implementing AI applications. The upcoming section focuses on the different attitudes towards AI Solutions in the medical domain, the barriers, and how others overcame these.

2.4.1 Opinions

Due to a study by “Impact of artificial intelligence on radiology: a EuroAIM survey among members of the European Society of Radiology” (2019) where 263 undergraduate medical students were questioned about their attitudes towards AI, most of the students did not think that algorithms will be able to make specific diagnoses in imaging examinations. But the students admitted that AI would automatically detect pathologies in imaging examinations and even indicate appropriate assessments. Furthermore, they found that knowledge and a basic understanding of the critical principles in deep learning and AI will be crucial for future generations of radiologists. Several papers pointed out that knowledge about AI workflows is essential in assessing these solutions, which leads us to the barriers AI faces when being implemented in clinical practice.

2.4.2 Barriers

Trust in medical technology is closely related to its anticipated utility. Disruptions of the current clinical workflow will inevitably face inertia, and if the perception emerges that there are untoward consequences of new technology, then the barrier to any similar technology will become next to insurmountable. Due to Yu and Kohane (2019) the most prominent issues of medical AI applications are: “The difficulty in identifying and updating a set of axioms to properly describe the environment for autonomous agents,” “trust in the performance of the AI program,” “amplifying biases presented in the historical data,” and “clinical workflow disruptions.” commonly agrees these barriers (Strohm et al. (2020)). “inconsistent technical performance,” “unstructured implementation processes,” “uncertain added value for clinical practice of AI applications,” and “large variance in acceptance and trust of direct and indirect adopters.” This research focuses on two of these problems: the lack of acceptance and trust and the insufficient implementation of AI solutions in the clinical workflow.

2.5 Visualization of AI and Machine Learning Processes in HCI

Learning how people interact with artificial intelligence-enabled machines and helping them understand the processes behind them will be crucial in the acceptance of AI solutions. As Lee (2018) discovered, users tend to attribute decisions made by algorithms to their efficiency and objectivity, which render them fit for mechanical tasks but unfit for “human tasks” that involve subjective judgments and emotional capabilities. Sundar (2020) have proposed a framework to study AI that may help researchers better investigate how people interact with artificial intelligence or Human-AI Interaction (HAI). Due to Sundar (2020) the framework identifies two paths — cues and actions — that AI developers can focus on to gain trust and improve user experience. Cues are signals that can trigger a range of mental and emotional responses from people.

2.5.1 Cues and Actions

Due to Sundar (2020) AI mediums can trigger cognitive heuristics by advertising their existence on the interface of the medium and providing interface indicators of, or about their output and their modus operandi. These heuristics, in turn, shape users’ psychological responses to the AI medium. As Kim (2016) discovered in his study of Internet of Things (IoT) devices, providing source cues by having each device communicate in a unique voice and designating a device as a specialist (rather than generalist) can positively affect user experience by triggering heuristics about social presence and expertise respectively. In this way, the cue route is primarily concerned with transparency and visible aspects of the AI system powering the medium. When the user engages with the AI system and provides their input, as in exerting control over an algorithm (the second aspect on AX), users can either choose to be guided by the machine and conform to its directives or control its customizing settings. Given the interactive nature of these systems, the action route is premised on the availability of volitional control for users (Sundar (2020)).

2.5.2 Transparent Design

O. and A. (n.d.) proposes the concept of Algorithmic Experience (AX) as an analytic framework for making user interactions with algorithms more explicit. For an ideal AX, users ought to be aware of how the algorithm functions and what it tracks to provide personalized services (Sundar (2020)).

Furthermore, they should be able to manage, corroborate and regulate its profiling, with the option of directing its future behaviors (for avoiding or emphasizing certain kinds of outcomes). The two critical features of AX—user awareness and user control—are the hallmarks of successful user experience with personalization services, as evidenced by several recent studies (e.g., Zhang and Sundar (2019); Chen and Sundar (2018)). It provides explanations about the how, why and what an algorithm can increase algorithmic transparency (Rader, Cotter, and Cho (2018)) whereas not doing so can lead to anxiety and trial-and-error reverse-engineering (Jhaver, Karpfen, and Antin (2018)).

2.5.3 Explainable Design

Aside from the identity of the source as AI, its visible attributes, such as system transparency and explainable AI (XAI), can trigger positive heuristics and thereby lead to better user engagement (Sundar (2020)). The context for the surge of XAI can be attributed to the fear of not understanding and controlling increasingly complex ML models. Explanations are often embraced as a cure for “black box” models to gain user trust and adoption. So a common pursuit is to produce interpretable, often simplified, descriptions of model logic to make an opaque ML model seen as transparent.

This effort is necessary but insufficient to deliver a satisfying user experience if we ignore users’ motivation for explanations. As I-8 put: “Explainability isn’t just telling me how you get there, but also, can you expand on what you just told me...explanation has its utility” (Sundar (2020)).

Due to a study by Liao, Gruen, and Miller (2020) the essential areas that showed the necessity of explanations would be: Input: Understanding training data for the AI model for comprehensive transparency of training data, especially the limitations. Output: the desire to understand the value of the AI system to evaluate the capability appropriately and better utilize the AI. How-global model: Informants recognized the importance of providing global explanations on how the AI made decisions to help users appropriately evaluate the system capabilities and build a mental model to better interact with or improve the system. Why, Why not–local prediction: Understanding a particular decision was often ranked at the top, and in user questions mentioned in all products (Liao et al. (2020)). Overall it seems like XAI research still struggles with a lack of understanding of real-world user needs for AI transparency and by far little consideration of what practitioners need to create explainable AI products.

2.5.4 Human Computer Collaboration

User actions afforded by AI interfaces can dictate user engagement and experience based on the extent to which they allow users to interact with the system, assure them of human agency, provide tangible benefits and offer avenues for mutual augmentation. This route is dictated by the nature of the collaboration between human users and AI systems, quite unlike the cue route, which is based on human perception of the manifestation of AI systems. It can also serve to mitigate perceptual concerns arising from the lack of algorithmic transparency. This is not always possible, either because of concerns being hacked or because the AI is a black box, based on self-learning, underlying patterns that are so complex that they maintain a mystery, even to the designer.

As such, the action route is likely to involve more effort than the cue route and determine user outcomes based on user engagement or involvement. If dual-process models in social psychology are any guide, trust in AI systems built via the action route is likely to be more robust than that via the cue route, based as it is on deeper involvement with the algorithms and perceived understanding of their functions, and therefore more resilient to change even under circumstances of failure.

Chapter 3

Discover and Define

3.1 Participatory Research Design

The main goal of this research is to gain insights into the work of radiologists, get to know the barriers in their clinical workflow, and discuss opportunities and their perception of Artificial Intelligence. These expert interviews should provide a baseline of knowledge to kick off the design experiment and the following development phase. The sessions focus on understanding current needs, fears, and expectations from real users. Additionally, a first assessment of the operating system deepcOS by deeps was part of the interviews. This participatory design approach should include real users from the start to provide a user-centered design that should be beneficial to the end-user.

The research aims to seek answers to the following sub-questions and address the following objectives of the thesis:

1. **SQ2:** What makes radiologists trust the output of a structured reporting artificial intelligence solution?
2. **OB2:** Understand how trust interrelated with artefact acceptance.

Instruments and Methods

Method	Research Question (RQ) / Sub-question (SQ) / Objectives (OB)	Comments
First Impression, Exploration, Show and Tell	<p>SQ1 How can we design a structured reporting artificial intelligence solution to support the clinical work-flow of radiologists?</p> <p>SQ2 What makes radiologists trust the output of a structured reporting artificial intelligence solution?</p>	<p>At first the users should get familiar with the software and explore it. Afterwards deepcOS is explained in depth by the researchers.</p> <p>The users should express themselves using the think-aloud technique. The researchers observed and collected notes during this phase.</p>
Human Computer Trust Scale (HCTS)	<p>SQ3 To what extent can trust be a key to leverage effective artificial intelligence structured reporting work-flow?</p>	<p>This should provide a first impression on the trust assessment in the current version of deepcOS</p>
System Usability Scale (SUS)	<p>SQ1 How can we design a structured reporting artificial intelligence solution to support the clinical work-flow of radiologists?</p>	<p>This should provide a first impression on the usability assessment of the current version of deepcOS</p>

Retrospective structured Interview	semi- SQ1 How can we design a structured reporting artificial intelligence solution to support the clinical work-flow of radiologists? SQ2 What makes radiologists trust the output of a structured reporting artificial intelligence solution?	The possibility to dive deeper into some observations or questions will provide valuable qualitative insights.
---	--	--

Prototype

We created a clickable, high-fidelity prototype of the current version of deepcOS for the expert interviews. This AI-driven diagnostic tool was originally designed by deepc. It is also available as a demo implementation. But to prevent distractions by implementation errors or other bugs, we decided to provide a clickable prototype limited to one specific use case. The prototype consisted of eight screens that were designed and implemented as a clickable version using the design tool figma¹.

Use Case Definition The prototype focuses on the workflow of reporting findings to explore how the future version of deepcOS may communicate AI model output more transparently. To give the user a clear starting and ending point, we separated the use case into several steps. These steps are described further in the following paragraph. The complete study plot can be viewed in Appendix A.

¹<https://www.figma.com/>

User Tasks Every time the participants entered a new screen, they should provide insights into their first impressions and thoughts, following the "think aloud" technique. The participants went through the following user scenario and tasks:

Imagine yourself sitting down in front of your working area and you start deepcOS to review the image set of the patient you have just screened.

Task 1:

1. Login to deepcOS
2. Select a patient from your list that has not been reviewed yet

Task 2:

1. Scroll through the images until an anomaly is shown
2. Now you want to document your findings according to the anomaly, please go to the area where you can add and edit findings.

Task 3:

1. Set a marker where you want to document your findings
2. You see that it is a medium localized hyperdense (a Intraparenchymal Hemorrhage) in the thalamus-right region
3. Please document your findings and finish the process

Survey

To measure the status quo and understand the users' perception of trust and usability of deepcOS, we used an online survey including the Human-Computer Trust Scale (HCTS) developed by Gulati et al. (2019) and the System Usability Scale (SUS) (?). The participants filled out the surveys after experiencing the prototype. We used Google forms to create the surveys. They were sent out to each expert during the session.

Semi-Structured Interview

To gain deeper insights into the individual workflows and perceptions of the related topics, we conducted a semi-structured interview after the participants completed the online survey. This method should provide the possibility to dive deeper into the following topics:

1. deepcOS prototype
2. Artificial Intelligence
3. Workflow
4. Modular/Structured Reporting
5. General Question

A full version of the interview can be found in German and English translation in the appendix. The semi-structured interview was closed by some demographic questions. These included the own perception of technical experience, age, gender, occupation in years, and highest educational degree.

3.2 Interviews

Since the interviews took place in Germany, the moderator interviewed all experts in the German language. To allow the readers of this thesis to interpret and understand the data provided, the analysis and results are presented in the English language. A complete study plot, in German and English translation, can be found in the appendix. To validate the results, the author is aware that an official translation may have been necessary to avoid possible threats to the study's validity. Due to the simplicity and the small number of data, a professional translator was not considered essential for the research.

Ethical Considerations Due to the ongoing COVID19 pandemic and to increase safety and flexibility for all connected parties, all interviews were conducted remotely, using Google Meet. To meet ethical research standards, participants were informed about the study's purpose and sent an invitation with an optional consent form to sign beforehand.

The consent form provided a declaration of data security for both sides (this form can be found in both languages in the appendix, too). Since participation in the Google Meeting required a google account, the author created an optional google account and sent it out beforehand, in case the experts preferred to participate anonymously. Before starting with the interviews, the participants were briefed regarding the purpose of their participation. They were also briefed regarding the methodological approach the sessions were following. The software was explained in general to the participants. They were informed that about the prototype being limited in its functionality to prevent frustration and confusion.

Participants

For this initial qualitative research, four participants from the medical domain were interviewed. All participants work in clinical environments of radiology. None of the participants were connected to this research before. Since this should provide first insights on how we will develop the experiments further, a small sample size should help discover the topic deeper.

Pilot Interview We planned the interview sessions for around about 45min-60min. We conducted a pilot test beforehand to test all technical conditions and go through the complete testing process. This way, we could eliminate the last mistakes in the clickable prototype or spelling errors in the surveys. The pilot participant took part in the meeting with the anonymous Google Account created for the participants. The technical process went flawless regarding participation in the interview and recording the session by the interviewer.

3.3 Data Analysis

3.3.1 Observation

While conducting the given tasks and browsing through the software, the participants were observed via screen sharing (3.1. After entering a new screen, the users had to describe what they see using the "think aloud" technique by Michael Häder (Einführung and ö Springer (n.d.)).

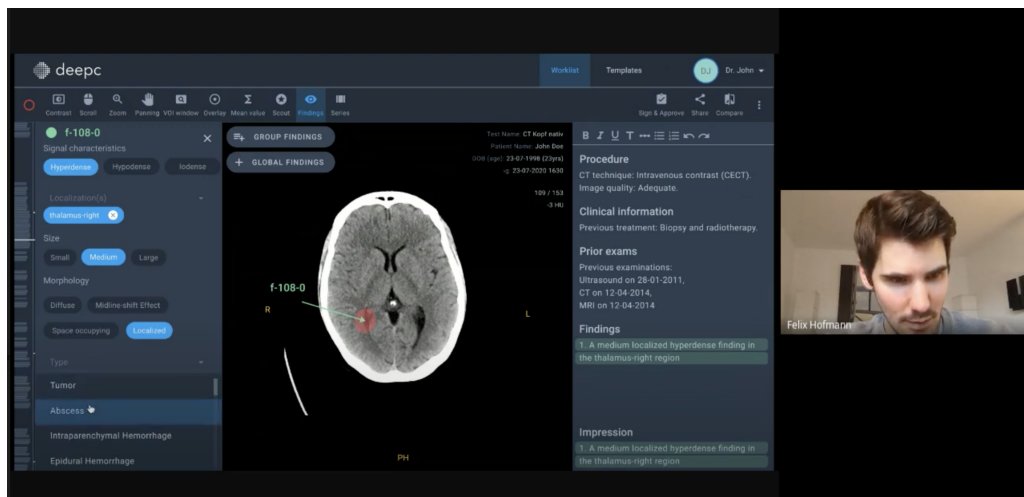


Figure 3.1: Expert Interview with Felix

It helped to understand their (unconscious) behavior patterns, first impression, handling of, and procedures for dealing with, the new software environment. All sessions were recorded and observation notes were collected.

Work-list The work-list was interpreted almost entirely correct by all experts. When the user started an examination, patients' names, date of birth, the purpose of review, MRI / CT, this information was visible. One or two experts had problems understanding what was meant by "Process," "Rating," "Classification," and "Anomaly Score." For example, the "Rating" was confused with the priority with which the user must examine a patient.

Image Viewer All experts correctly described what was displayed by the software: Axial CT of the head in the soft tissue window. Not all users could interpret the vertical bar on the left edge of the screen. Among other things, the users mentioned that there is no axis labeling here. The display of the slices of the CT was mistaken for a view of pixel values. Also, they took the length of the bars for the anomaly score. Two out of four experts would hide the mask in which the information from the AI is displayed again to see whether they could notice an anomaly on their own. The experts recognized many functionalities intuitively, such as scrolling up and down, changing the opacity and switching between layers.

Findings All experts correctly recognized the report area on the right side of the screen. However, the function to expand the findings with their documentation was requested, especially in this area.

None of the experts found the button to reach the area to document findings intuitively. It was more likely to see it as a function to jump back and forth between the different markings of the AI.

One of the users mentioned that the upper bar was only seen in connection with the image material, i.e., manipulating the image view and not documenting findings. An expert would have suspected the documentation of results under "Templates." In this view, the experts also wanted other tools, e.g., the painting of arrows, for measuring the distances. One of the experts mentioned that in the status quo, all markings are later deleted again. If they exist for more than four to five days in the DICOM format, they are burned into the image. Also, drawing an arrow into the picture, the expert would do it outside the image content. This led to the realization that the markings must generally be implemented as an overlaying mask so the original scan is not being damaged (see figure 3.1).

3.3.2 Survey

System Usability Scale (SUS) The average score of the SUS questionnaire between all four participants (we excluded the results of the pilot-test from these results) is 69,4. It is considered above-the-average within the industry. The average score is 68. Only one participant evaluated the prototype on the grade of 60. In the following diagram (see figure 3.2) the results from all four participants are visualized:

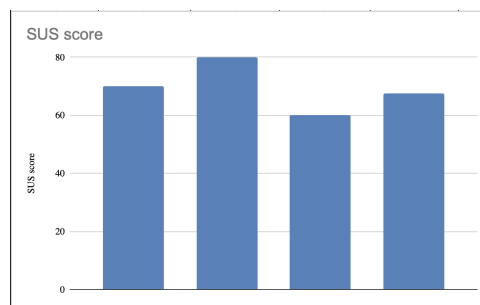


Figure 3.2: SUS Score of the experts interview

Human Computer Trust Scale (HTCS) We chose the Human-Computer Trust Scale (Appendix: User Study Plot: Expert Interviews) to measure the baseline of trust in deepcOS. After interacting with the prototype, we asked the participants to reflect their assessment of trust in the provided software. Two experts rated the software with 84% and 83%. It can be considered above average and acceptable without any need for adaptations or redesigns. One expert rated the software with 74%, which is regarded as slightly below average but still satisfactory, only suggesting marginal changes.

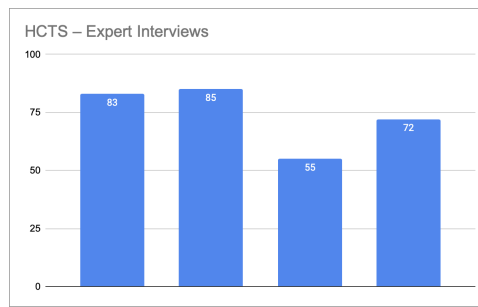


Figure 3.3: HCTS Score of the experts interview

One expert rated the software with 55%, which can be seen below average and needs significant adjustments (see figure 3.3). Therefore the average for deepcOS is 74%, which points out that view adjustments need to be done to enhance trust in the current version of deepcOS.

3.3.3 Interviews

We chose to create a codebook to compare the gathered qualitative data collected by the semi-structured interviews. Having the theories in mind that this thesis focuses on, priority coding relied on categories based on previous research. By evolving the codebook, we enhanced the previously selected categories with new ones based on interesting concepts or ideas that came up during the interview. Thus the general approach of the codebook can be seen as a mixture of priority and emergent coding. In the first stage of creating the codebook, we collected the experts' answers in a table, sorted by experts (e.g., E1), question, and question category (e.g., Artificial Intelligence). In the second stage, we added the previously defined coding categories with colors. The third stage included coding the text. Therefore we colored the text passages according to categories (see figure 3.4). Additionally, we added connected colored codes on the side of the text passage. The codebook can be found in the appendix. Due to the limited scale of this research, the reliability of this codebook was not verified by additional coders.

Question	Text extracts from semi-structured interview	Coding
deepcOS	Design, AI, Usability, Features, Support, Requirements, Trust	
What did you like about the software?	E1: "New look for findings; wall is broken, linking of text data with localization in the picture; In the process still filled with AI. In the picture documentation, structured findings" E2: "Stylish, pleasant to use from the design experience, simple and not unnecessarily complex. Extremely well resolved. The viewer is also good" E3: "Design is beautiful, clean, clear, if you use it more often then you know where something is, no major training required, relatively self-explanatory,"	
How can the software support you with your individual requirements?	E1: "Integrated in the workflow, seamless integration, to be incorporated individually through the template, can be adapted." E2: "That is pointed out again to other things that I might have forgotten. Statistically it happens that something is overlooked. The software should support there. The paperwork should reduce by clicking through. Connection with voice control would be important. Periodization, there are signs of what is a serious pathology - what can be put back - that is helpful" E3: "Got to play around with it, you have to get a feel for how trustworthy it is. Look at studies, system declared as normal - look again yourself to confirm this. Am I the only one using it? Do you have previous experience? If I am the only one who uses it and is new, then I would look at it a lot longer"	
What looked strange or didn't you enjoy going through the software?	E1: "Findings to open the templates, intuitively thought to jump from finding to finding." E2: "That I would have clicked in the report window first would have been good if both ways work. Markings or annotations, arrows, circles - the function should be more prominent." E3: "What bothered me a little was that the marking of the pathology, it bothers me when something is in the picture. I would like to click away to see it for myself and it would be important to me that the marking is removed quickly. I would prefer that the radiologist is sorted first in the working list. You couldn't scroll - but I think it was due to the prototype. 1 computer screen for the text / findings + 2 workstation screens on which we can view the images. Findings are right next to the picture. Findings are on the left and images are in front of me or on the right. It is better to have the findings separately and photos alongside. CT images are series, you can use different contrasts, several series side by side."	
Which functions did you miss that are not implemented in deepcOS?	E1: "Two diagnostic monitors - multi-monitor operation. Compare function as centrally important. Axial slice guidance for the findings, often 3D images that are relevant and important." E2: "Emphasize even more with pictograms e.g. the drawing tools. In the case of findings, a pen next to it so that you can then fill in the structured report. As a radiologist you have several screens, it would be nice if it were designed so that this matrix goes through on the same screen as the image. The picture could be smaller - split over two screens - or different windows arranged next to each other. Invert the colors for dark and light designs." E3: "Besides fading out the overlay, you have to be able to create distances, you have to be able to draw reach of interest. Zooming, scrolling, in the Pacs CTs slices can be recorded and can be reconstructed so that other planes can be seen."	
AI	Positive, negative, transparency	
Now you had the opportunity to check the results yourself. How would you feel if the AI were to fill in the parts you controlled?	E1: "Super should be done. Stupid manifestation of multiple sclerosis - software transparently marks where it got the findings from." E2: "This is a helpful development for me. These tools are important so that the radiologists can keep up at all. Gives radiologists more time to care for patients and research. Contact with patients is being neglected. When the machine takes over tasks and the radiologist maps the context and initiates quantitative analyzes, I feel very comfortable with it and look forward to using these tools further" E3: "Depends on the experience with the system. I have to be able to choose it. I would not care - I would change it or issue it. If I am very experienced it would be a relief. AI remain problem with the whole thing is - how much you allow yourself to be influenced. How events do I look at the other slice where the AI has not"	

Figure 3.4: Example code book

deepcOS In the category "deepcOS" that focuses on the prototype itself, the following codes were mainly used. **Design and Usability, Features, Requirements, and Trust.** Regarding **Design and Usability** the prototype was evaluated in general very positive: "The design is beautifully clean, clear;" "Extremely well resolved," "The design experience is simple and not unnecessarily complex". There were also some remarks to improve the usability, e.g., "I would have clicked in the report window. First, it would have been good if both ways work" or "Emphasizing more with pictograms, e.g., drawing tools. In the case of findings, a pen next to it so you fill in the structured report." The experts mentioned some **Features** they would like to see in the future, like "Axial slice guidance for the findings in 3D" or "Expanding the drawing tools, creating distances, circles". In case of **Requirements** the experts stated that it would be important that "it is seamlessly integrated into the workflow", or "The software points out to me what I might have forgotten". One of the experts stated that it would be essential for her to remove all remarks very quickly to look over them again. She also stated concerning **Trust** that she "got to play around with it, to get a feeling how trustworthy it is."

Artificial Intelligence In the category "Artificial Intelligence" the codes **positive, negative, transparency, trust, acceptance** and **support** were mostly prominent. Two of the four experts had strong **positive** feelings towards the integration of AI in their workflow. "This is a helpful development for

me.” *”It gives radiologists more time to care for patients and research,” ”I feel comfortable with it and look forward to using these tools further”*. One of the experts expressed some **negative** feelings using AI services *”How much do you allow yourself to be influenced by it? How exactly do I look at the other slices where the AI has not found any findings”*. When the experts were asked what it would take for them to trust the outcome of the AI prediction, they all mentioned **transparency**: *”software transparency, marks where it got the findings. Classification tasks, why does the AI decide that this is a negative prognosis, e.g., using regression models”*, *”Solving the black box problem would help, but in the future, it could be too complicated for humans to understand”*, *”Where does the data come from?”*. Other important factors for them to enhance **trust** would be to have gone through *”100 examinations myself which have same or similar results”* or to provide a set of studies that verified the AI. Additionally, an *”index how sure the AI is”* would be helpful, *”it depends on the performance - if it is 95 % certain, then I don’t need to do anything.”*. One of the experts stated that *”If the radiologists had their freedom for as long as possible, that would improve acceptance”*.

Workflow In the category *”Workflow”* the codes **picture**, **features**, **time** were the most prominent. The experts described two different workflows that are interesting for the development of this thesis and can be seen as the main basis for the upcoming personas. Two experts described the workflows from assisting radiologists that go through the pictures, writing reports, and walking through these reports with their supervising doctor later on. One participant described the other workflow by one of the supervising radiologists going through the reports and findings of the others. He said that for him *”each assisting radiologists can be seen as a tiny AI that needs to be checked or trusted.”* Two of the experts mentioned that the most **time** consuming task is to look at the **pictures**: *”Most of the time I’m looking at pictures,” ”Take a close look at the investigation and research”*. For one expert, the *”writing a report”* is the most time-consuming. There were some **features** the experts would like to see in the future to ease up their workflow: *”comparison between two studies, having a tumor patient, to see if the tumor has enlarged or spread. There should be a focus on what has changed since the last investigation.”* or *”The pictures perfectly hung up in the PACS system, depending on the question of course.”*.

Modular Reporting In the category of **modular reporting** there were two main codes supported: **positive aspects**, **negative aspects**. In general, one can say that there is neither a shift to a positive attitude according to modular reporting than to a negative one.


All experts had points pro and against it. As positive aspects, they mentioned that *"one can interpret findings more completely and more clearly," "standardization is important," "time-saving in the future," "helps training the AI," "can be an advantage if you are not yet that experienced"*. As possible downsides, the experts stated that *"you always have to click your way through everything," "it is impossible to depict all cases," "it could take a long time to go through the list, having to look away from the picture," "Dictating would be faster and I can still see the picture," "if you rely too much on it you can forget something"*.

3.3.4 Summary and Resulting Personas

We observed that the design and the usability of deepcOS are already on a reasonable level. The triangulation of investigation can follow this, having the positive results of the SUS questionnaire (68,4 score) and the outcome of the semi-structured interview. The current score of trust 74% leads to the conclusion that the software needs minor improvements to increase the level of trust.

The expert interviews now provided valuable input for the development of the prototype, and this way, we found an answer to *SQ2 – What makes radiologists trust the output of a structured reporting artificial intelligence solution?* as well as *OB II – Understand how trust interrelated with artifact acceptance*

The codebook helped collect and compare the data, identifying risks, opportunities, features, needs, and expectations. The following personas should provide the resulting defined goals from the user perspective (see figure 3.5 and 3.6).



"If I'm not multitasking everything, I'm actually less efficient."

AGE 37
JOB TITLE Senior Radiologist
EXPERIENCE 2 years in the Job
ENVIRONMENT University Clinic
LOCATION Munich, Germany

TECHNICAL EXPERIENCED

CURIOUS **EMPATHETIC**

USER PERSONA

Sarah

DIAGNOSTIC RADIOLOGIST

ABOUT

Sarah is a young dynamic senior radiologist. She is a real multitasker. Obviously, she needs to focus on the task at hand at any given moment. But overall in life, she needs to be working on multiple things. Sarah currently is a practicing radiologist at the MRA (Clinic Rechts der Isa in Munich, Germany), but previously spent a lot of time in academics. Thus she likes to get to know and try out innovative solutions that could improve the status quo of her daily workflow. Due to the great workload she seeks for solutions that help her to find time for her patients again and for research activities.

GOALS

- Provide feedback to assisting colleagues faster.
- More frequent exchange with colleagues about cases and anomalies, expanding research opportunities.
- Having all the information needed to do the job

PAIN POINTS

- The communication of different devices and therefore patient data does not work.
- Not having enough time to go through all cases properly in a day


NEEDS

- Comparing functionality for different studies
- Seamless integration of new solutions into her personal workflow

SCENARIOS

- Review studies
- Reviewing assisting colleagues cases and notes
- Researching patient history

Figure 3.5: Persona: Senior Diagnostic Radiologist



"I like to connect with patients and to get involved more deeply into research"

AGE 30
JOB TITLE Junior Radiologist
EXPERIENCE 2 years in the Job
ENVIRONMENT University Clinic
LOCATION Munich, Germany

TECHNICAL EXPERIENCED

CAREFUL **EMPATHETIC**

INQUISITIVE

USER PERSONA

Felix

ASSISTING RADIOLOGIST

ABOUT

Felix has been an assisting radiologist for two years at the MRI in Munich. He works there more than 50 hours a week and after that he can hardly leave work be, because of the huge amount of work he sometimes can't finish in a day. To overcome this problem he likes to get involved with startups to work on a solution that could help in the near future. For him all new technology needs to be studied and validated carefully over years, so for him it is important to understand the processes and participate in studies and development of solutions to estimate the feasibility.

GOALS

- Handle the amount of cases to scan in a day
- Provide flawless reports
- Having all the information needed to do the job

PAIN POINTS

- The huge workload of analyzing cases properly in a day
- There is not enough time to connect with patients or involve in research

NEEDS

- Security that the findings are valid
- Support in analyzing findings
- Support in research

SCENARIOS

- Analyze cases, document findings and write reports
- Researching patient history

Figure 3.6: Persona: Junior Diagnostic Radiologist

Chapter 4

Design Development

Due to the preliminary expert interviews and following analysis, some system adjustments are necessary to design for transparency and improve the workflow (e.g., where and how to document findings using drawing tools). One more significant challenge might be the distribution of the interface on multiple screens. Due to the limitation of this thesis, this feature needs to be prioritized less. The final study needs to be limited to one screen interaction due to the ongoing COVID19 pandemic. The author recommends working on this feature in future related work. There are three possibilities to improve the level of trust using the system based on the outcome of the expert interviews:

1. Design for transparency to solve the black box problem of AIs
(focusing on visual and content cues)
2. Provide studies about the system (focusing on social cues)
3. Provide a four eye support solution (focusing on social cues)

There should be a feature considered regarding the workflow integration that lets radiologists compare the studies quickly and efficiently to save research time. One significant barrier that needs to be taken into account in future work is combining patient information from different sources. There are still many different systems in the field (e.g., KISS and PACS) that do not communicate with each other. This information needs to be brought together. The medical community is already aware of this problem, according to the previously conducted literature review. Solving this problem of different, sometimes also unstable, systems and information sources would also help speed up and ease up the

individual workflows. For modular reporting, a solution needs to easily browse through the labels/suggestions without losing focus on the diagnostic image itself. A possible solution here could be an auto-complete text input combined with creating new labels for the database.

For this prototype's development phase, the goal "Design for transparency to solve the black box problem of AIs" was chosen.

Use Case Definition To create the basis for an honest comparison in the summative study, we chose the same use cases of the current clickable version of deepcOS. The users should log in to the system, select a patient who has not been reviewed yet and document his*her findings where the AI has detected anomalies. This use case visualized the main workflow and was redesigned in a second clickable prototype.

4.1 Design Adaptions

We did not change the general design language itself in this version. During the experts' interviews, it was clear that it was seen as professional and stylish and therefore implemented sufficiently. The adaptations of the prototype, focus on three main changes and were designed based on an adapted version of the design for trust checklist by Wang (2009):

1. The workflow was redesigned with the goal to increase the usability.
2. A trust design check-list based on Wang (2009) was used as a guidance during the design process. The check-list includes visual, content and social cues that improves the assessment of trust in the software.
3. To design for transparency cues regarding the artificial intelligence were implemented into the software. These cues should help the users to understand the AI processes and break up the black box.

We created all design adaptations and clickable prototypes for Version A and Version B with Figma. Since Figma offers easy HTML prototype integration in the "loop11" testing software, it seemed sufficient to use it as a design tool.

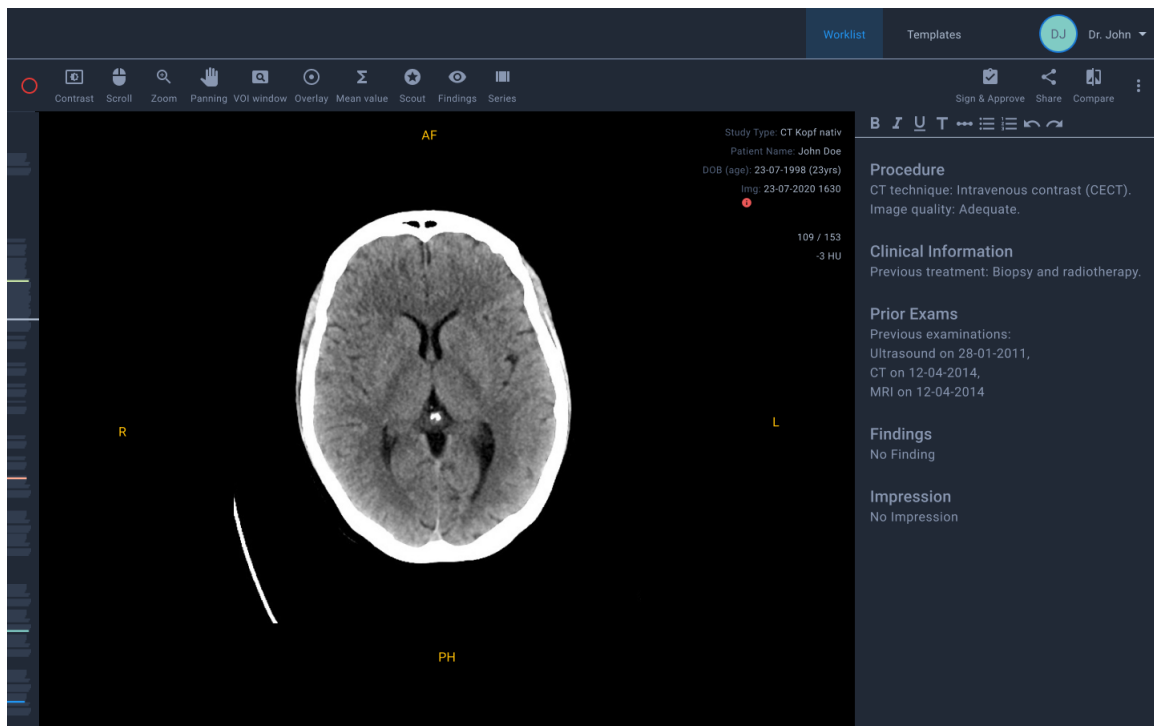


Figure 4.1: Version A of Image Viewer

4.1.1 Usability and Workflow Adaptions

Work list We rearranged the layout of the worklist regarding the importance of the provided data. Due to the experts, the essential information is the name of the patient. Thus, we reordered the patients' names to the first slots. The AI rating, along with the anomalies score, was reordered to the back of the list. Due to the small focus of the study, the density of expert information provided in the prototype was never reduced or changed. Only in the case of AI representation was information added to provide transparent workflows. Therefore the author recommends investigating, in a follow-up study, the necessary details that need to be displayed in the worklist (e.g., patient ID or date of birth). Additionally, we changed the representation of the AI rating to a visualization consisting of three compounds: AI icon with the result (healthy, abnormal, unavailable), overall rating, and doctor icon (healthy, irregular, unavailable)(see figure 4.3 and 4.4).

Layout The layout was broken up into specific sections (Header, Slices, Image Viewer, and Report) to support a responsive adaption in the future (figure 4.1 and 4.2). As mentioned before, the experts would like to have the possibility of a modular interface that can be distributed on multiple screens. Also, for additional guidance, the modular sections were clustered in specific areas and labeled to improve the overall orientation.

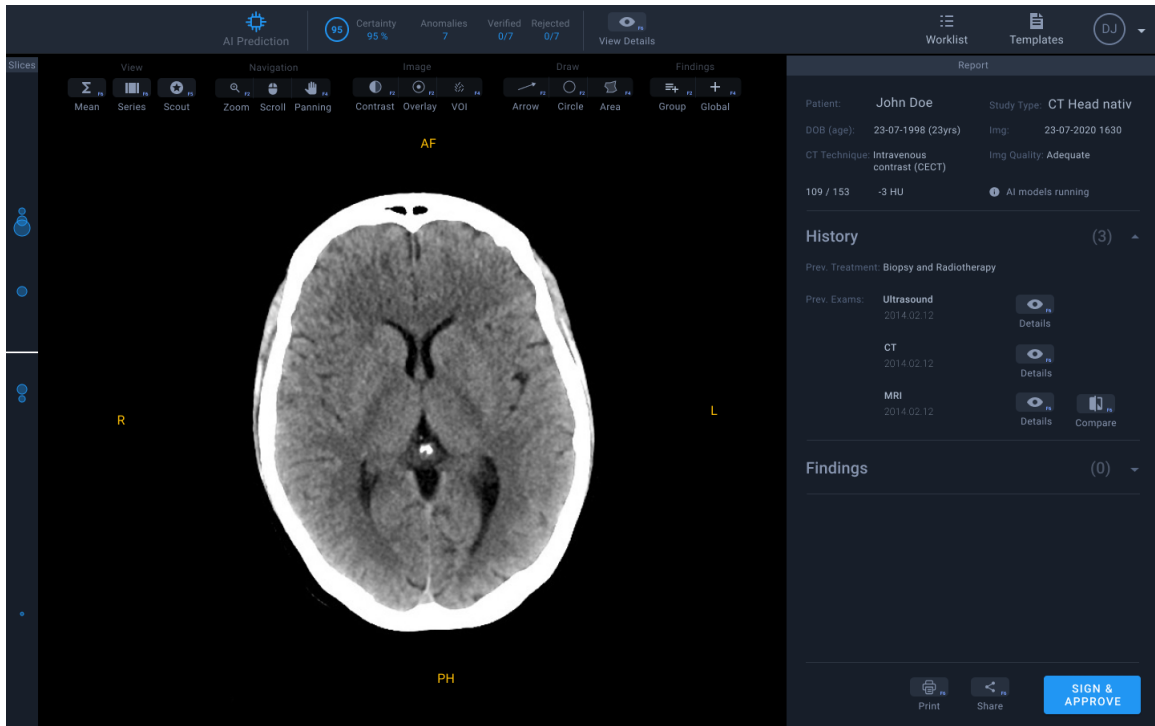


Figure 4.2: Version B of Image Viewer

The screenshot shows a 'Worklist' table with the following columns: Process, Patient Name, Classification, Study Date, Study Time, DOB (Age), Examination, Doctor Name, AI Rating, Anomaly score, and Reports. The table contains 12 rows of patient data. The AI Rating column uses a gear icon with a checkmark or an 'X' to indicate the status of the AI prediction. The Anomaly score column shows numerical values ranging from 0.012 to 0.975. A footer note states: 'For investigation and research purposes only. This is not a certified medical device.'

Process	Patient Name	Classification	Study Date	Study Time	DOB (Age)	Examination	Doctor Name	AI Rating	Anomaly score	Reports
<input type="checkbox"/>	Jane Doe	-	26.08.2020	9:00 AM	12.04.1998 (22)	BRAIN	Aaron Young	⚙️ ❌	0.012	📄
<input checked="" type="checkbox"/>	Martin Merces	-	26.08.2020		12.01.1987 (33)	MR, Brain, KM	Aaron Young	⚙️ ❌	0.975	📄
<input checked="" type="checkbox"/>	Cindy Sherman	-	26.08.2020	8:12 AM	31.08.1965 (55)	MR, Brain	Aaron Young	⚙️ ❌	0.452	📄
<input checked="" type="checkbox"/>	Joseph Gonzalez	-	26.08.2020		23.11.1931 (88)	CT, Brain, nativ	Aaron Young	⚙️ ❌	0.895	📄
<input checked="" type="checkbox"/>	Nikol Ricci	Normal	26.08.2020	6:53 AM	26.05.1983 (37)	CT, Shoulder, nativ	Aaron Young	⚙️ ✅	0.098	📄
<input checked="" type="checkbox"/>	Judith Williams	Normal	26.08.2020	5:18 AM	1.06.1997 (23)	CT, Abdomen, KM	Aaron Young	⚙️ ✅	0.036	📄
<input checked="" type="checkbox"/>	Franz Ferdinand	Metastasis, Subacute Stroke	26.08.2020	3:43 AM	01.10.1978 (41)	CT, Brain, nativ	Aaron Young	⚙️ ❌	0.956	📄
<input checked="" type="checkbox"/>	George Fields	Epidural Haemorrhage	25.08.2020	11:38 PM	28.03.1943 (77)	MRI, Brain	Aaron Young	⚙️ ✅	0.312	📄
<input checked="" type="checkbox"/>	Lindsey Stroud	-	25.08.2020	11:12 PM	19.07.1964 (56)	MRI, Knee, KM	Aaron Young	⚙️ ❌	0.859	📄
<input checked="" type="checkbox"/>	Rebecca Moore	Normal	25.08.2020	10:38 PM	30.04.1937 (83)	MRI, Thorax	Aaron Young	⚙️ ✅	0.011	📄
<input checked="" type="checkbox"/>	Jones Dermot	Normal	25.08.2020	10:02 PM	23.07.1955 (65)	CT, Brain, nativ	Aaron Young	⚙️ ✅	0.005	📄
<input checked="" type="checkbox"/>	Martin Merces	Subarachnoid Haemorrhage	25.08.2020	10:01 PM	28.02.1978 (42)	MRI, Brain	Aaron Young	⚙️ ✅	0.345	📄

Figure 4.3: New design of the work list

The comparison shows two designs for explaining an AI score. The left design shows a simple 'Observation' and 'Anomaly' column with a score of 0.798. The right design provides a detailed breakdown of the score based on radiologist and AI model results.

Observation	Anomaly
○	0.124
○	0.048
●	0.798

Radiologist	Overall Result	AI Model	AI Model Result
radiologist result healthy	healthy	AI Model Result healthy	healthy
radiologist result abnormal	abnormal	AI Model Result abnormal	abnormal
radiologist has not looked at	not assessed	AI Model Result undetermined	undetermined
		AI Model Result no data	no data

Figure 4.4: Comparison and explanation of AI score for original (left) and new design (right)

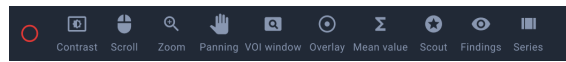


Figure 4.5: Version A of the global Toolbar

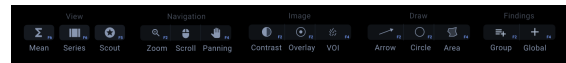
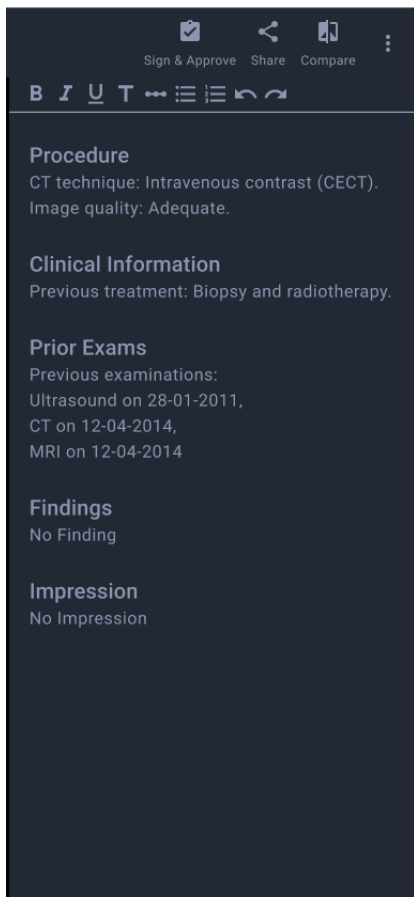


Figure 4.6: Version B of the global Toolbar

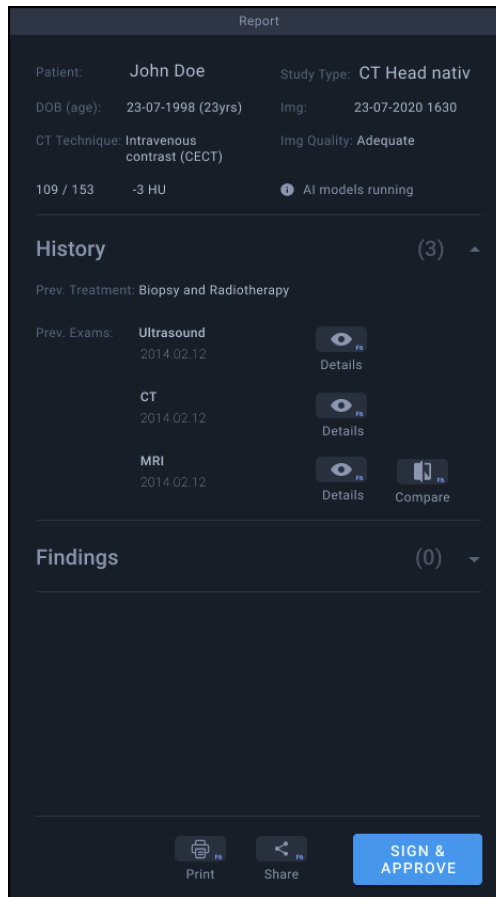
The user should be guided in reading directly from the slices to the selected image and the reporting section. The patient information incorporated in the image view before, now moved into the report section to collect all related information in one space (read about more adaptations in the "Report Findings" paragraph).

Toolbar Based on feedback in the expert interviews, we made some adaptation to the upper toolbar in the image viewer (see figure 4.5 and 4.6). The upper bar was only seen in connection with the image material, i.e., manipulating the image view and not documenting findings. Also, the need for drawing tools was an essential feature for the experts. Therefore the toolbar was separated from the header and included in the image area. It should clarify to the user that these tools are related to the image only. Additionally, the tools were clustered in different categories to improve the overview of the options and their connections. We incorporated the AI validation status into the header in the new version, using a red color scheme (see section AI representation). We also excluded the button "Findings" from the toolbar and integrated it into the slices bar on the left side and the report section on the right side to support the affiliation.

Report We adapted the report section in multiple ways (see figure 4.7). As mentioned before, we combined the latest patient screening data at the top, together with the "Procedure." The following section, "History," helps the user find all information regarding prior exams or "Clinical Information" Data. The user could open available documentation or even compare it if they used the same image modality. A small number at the top should communicate to the user how many prior examinations are available. The sections can be opened and closed to focus on essential data that are relevant to the user at the moment. We moved the writing tools to the "Findings" section below to provide the necessary connection. The "Share" and "Sign and Approve" buttons were transferred underneath the "Findings" section to give a chronological order where the user goes through the process from top to bottom until he*she signs or shares the findings in the end.



(a) Version A



(b) Version B

Figure 4.7: Report Section Version A and B

Additionally, a "Print" button should provide fast access if the users need to print the report.

Reporting Findings Workflow The changes of the reporting workflow we made are significantly different from the original version (see figures 4.8 and 4.9). First of all, the reporting was condensed entirely in the "Report" section at the right. To provide the user a consistent workflow in reading direction, the user would first select a slice he*she wants to look at, then examines the image and continue reading to write the report finally. This new workflow should help the user to concentrate on one specific task without jumping too much back and forth between the screen elements. "Group" and "Global Findings" were integrated into the toolbar to maintain consistency. Furthermore, there is no need to click the "Findings" button anymore. As mentioned before, we incorporated the documentation of the findings in the reporting. The user clicks into the findings text area to start writing a report. One big difference to the user is that they do not click together the parameters from a specific selection. The user starts to write a report, and in that time, the artificial intelligence provides suggestions for auto-completion.

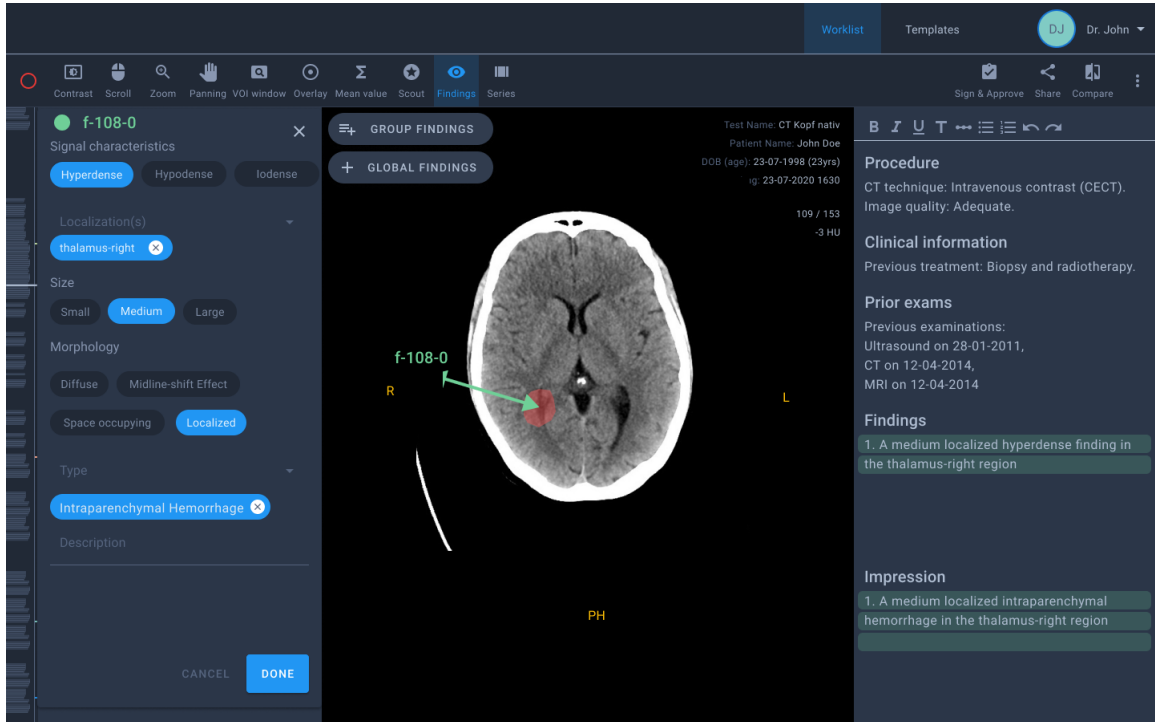


Figure 4.8: Version A of the documentation workflow

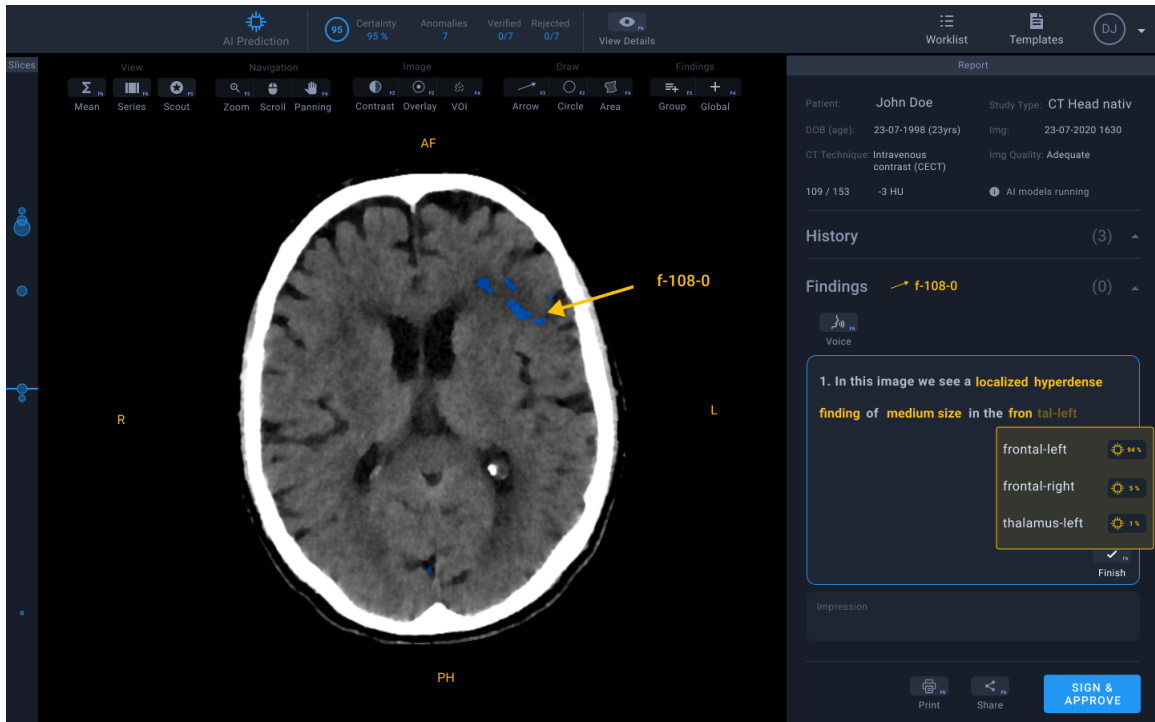


Figure 4.9: Version B of the documentation workflow

We chose this workflow because this is the current way radiologists document their findings. It combines a known strategy with artificial intelligence advancements in a familiar way they learn, e.g., from mobile text editors. Instead of selecting the "Impression" of the finding from a specified list, the user now gets a prediction from the artificial intelligence that can be accepted or declined, with further possibilities to write an own impression (more about the integration of AI can be read in the section "4.1.3 Design for Transparency").

4.1.2 Design for Trust

As mentioned before, due to the focus on workflow integration and design for transparency to solve the black box problem of AIs, most adoptions are related to visual and content cues. Since sharing and external sources are not yet part of the standard workflow, it was excluded for this design iteration. Nevertheless, the author sees the relevance and the importance of these opportunities and recommends including the social cues further in the subsequent design iteration.

Visual Cues

In the following table the visual cues are listed with an explanation if they were sufficiently existent or adapted.

Cue	Existent Yes / No	Comments
Professional design	Yes	-
Color scheme to suit the product/service	Yes	The color scheme was improved in the second version regarding AI color representation and call to action color scheme. (see figure 4.9)
Nice and legible fonts	Yes	The software uses the Roboto font which was developed for display representation. Thus the font was not adapted.
High-quality (and authentic) images and visuals	Yes	The visualisation of the imaging was adapted in the second version regarding anomaly detection in the brain scan using a realistic highlighting. (see figure 4.9)
Good on-site search	Yes	The image search was adapted in the second version regarding search for anomalies in the set of slices. (see figure 4.12)

Easy-to-use navigation	Yes	The navigation was adapted in the second version regarding tools and integrated reporting. (see figure 4.6) and 4.7
Clear anchor text and microcopy	Yes	The connection to the AI was improved in the second version using clear labeling and auto completion in the report section and creating a new layout for patient data. (see figure 4.7)

Content Cues

In the following table the content cues are listed with an explanation if they were sufficiently existent or adapted.

Cue	Existent Yes / No	Comments
No technical problems (broken links, missing pictures or pages, etc.)	Yes	The prototype matches the requirements of a flawless walk-through. We could not exclude technical problems entirely due to the remote testing application.
Brand-promoting information (logo, slogan)	Yes	A new logo was designed to hide the name of deepc due to content con-fidelity agreements. (see figure 4.11

Company information	No	Since we set the focus of the prototype on workflow improvements, we did not include the connecting information of the developing company in the current prototype. It could be included e.g., after clicking on the logo.
Contact information	No	Since we set the focus of the prototype on workflow improvements, we did not include the connecting contact information of the developing company in the current prototype. It could be included e.g., after clicking on the logo.
Useful (expert-level) content	Yes	The connection and integration of the AI combined with the workflow integration was adapted in the second version. (see figure 4.7)
Good grammar and minimized jargon	Yes	The prototype uses a clear and self explanatory language in case of icon language, labels, texts, headlines and AI communication (see figure 4.6)
External links (sources)	No	The connection to external sources was no focus in this iteration of the prototype. But in future developments, studies and information from external sources should be integrated.

Helpful FAQs	No	Due to the focus on the main workflow, we did not integrate FAQs or help in this iteration of the prototype. It needs to be included, e.g., in the settings in future developments.
---------------------	----	---

We have excluded the following categories due to missing relevance to the application: "Up-to-date blog", "Changelog", "Developer case studies", "Clear Terms of Service (Developer oriented that differ from company's own ToS)" and "Provide clean and working code examples."

Social Cues

In the following table the content cues are listed with an explanation if they were sufficiently existent or adapted.

Cue	Existent Yes / No	Comments
Easy access to customer service (e.g., contact form)	No	Since the prototype focuses on workflow improvements, the connecting information of the developing company was not included in the current prototype. It could be included, e.g., after clicking on the logo.
Developer Testimonials from successful businesses that built an integration	No	The connection to external sources was no focus in this iteration of the prototype. But in future developments should include studies and information from external sources.

Reviews from influencers and notable developers	No	The connection to external sources was no focus in this iteration of the prototype. But in future developments should include studies and information from external sources.
Social presence (Developer community, social media)	Yes	The users have the possibilities to share their findings with the community. (see figure 4.11)



Figure 4.10: Version A of the documentation workflow

4.1.3 Design for Transparency

To design for transparency, we implemented cues concerning artificial intelligence into the software. These cues should help the users to understand the AI processes and break up the black box.

AI Representation

To increase transparency, in connection with artificial intelligence, it was seen necessary to increase the representation of AI in the software in the first place. An additional highlight color, blue, was included to communicate data connected to the algorithm to create a clear connection. In the original version, the AI was only represented slightly. In the worklist (see figure 4.3) it provided a rating and an anomaly score. In the image viewer showing, the AI was visible with the rating only. Due to the importance of AI, we decided to:

1. Redesign the header integrating several information about status and validity of the algorithm (see figure 4.2).
2. the "Slices" bar on the left side was adapted to increase the connection to the AI
3. The workflow was adapted to provide a dialogue between human and machine

Status of Validity In the new design, the information "AI prediction" introduces the following parameter to connect to the AI algorithm. The data is connected to the currently selected set of images. Next to the certainty, the number of anomalies found by the algorithm is shown. This information correlates with the number of cues provided in the left bar where the user can scroll through the slices. Since the AI made some predictions about anomalies, a score shows the user how certain it is about its findings. The certainty increases /decreases with the number of validations or rejections the users provide at the end of each documentation. Due to the expert interviews, the users need to know the status of the algorithm's validity. The level of training and, therefore, the level of certainty can help here to provide information about the quality of the output data.

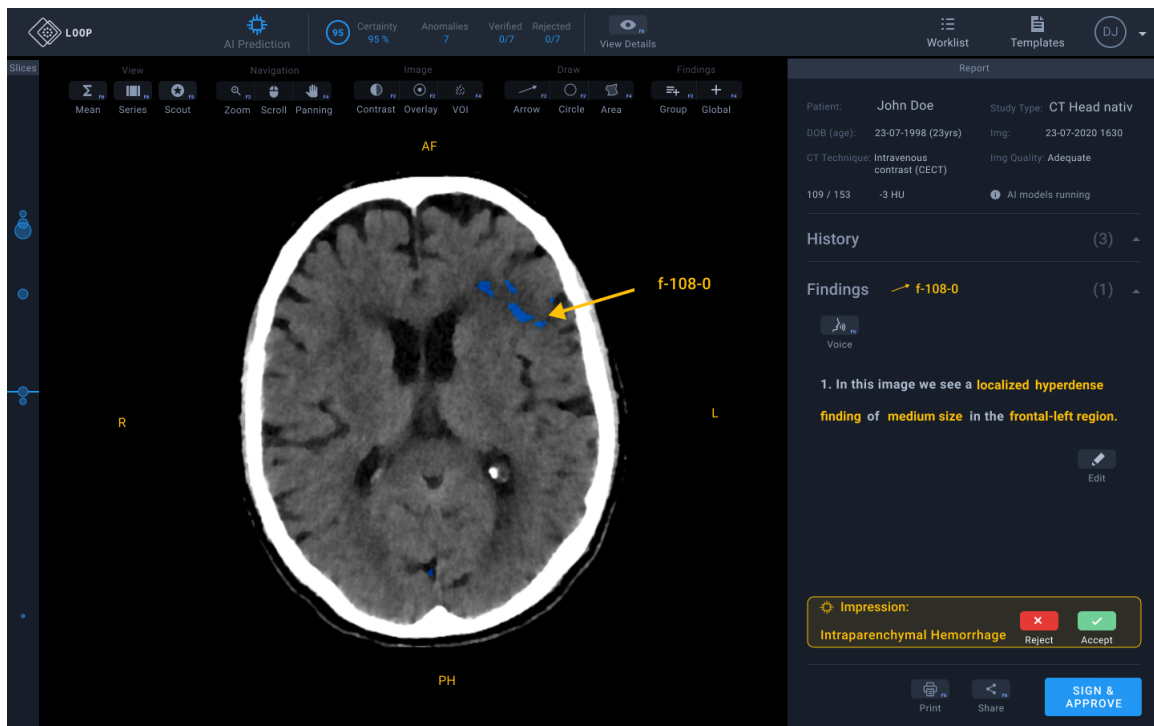


Figure 4.11: Impression, done by the AI

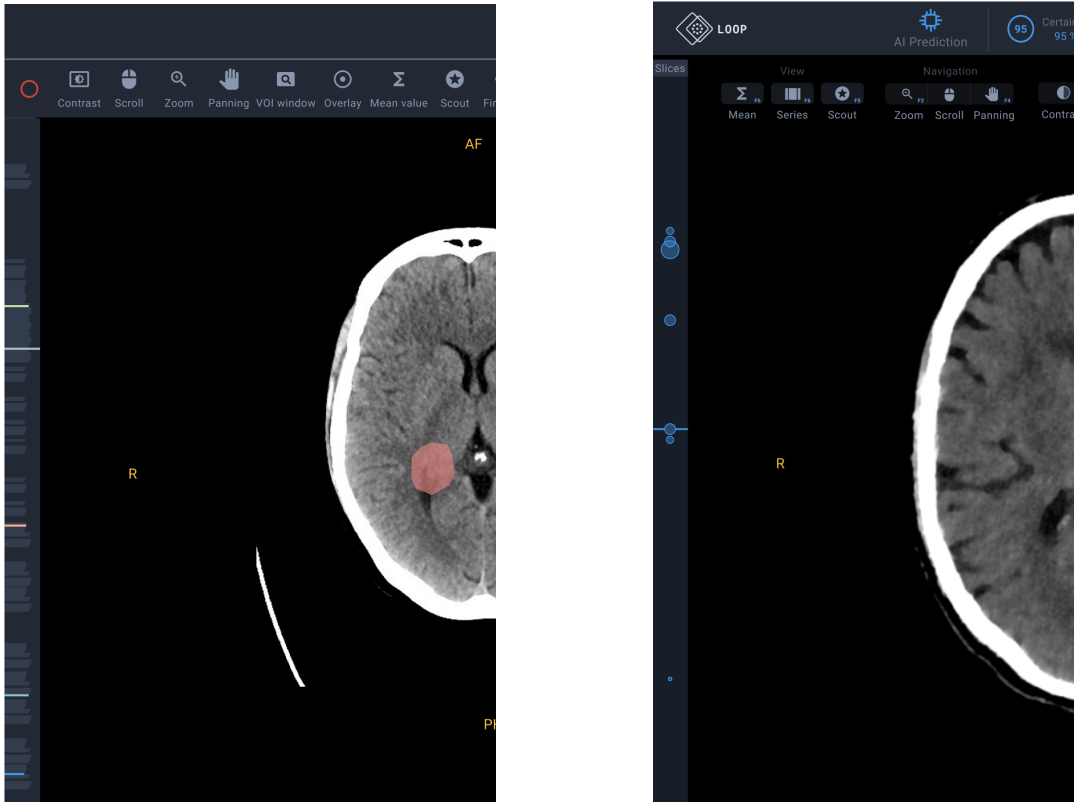
The estimation about output quality can be additionally supported by providing information about the source under "View Details." The first idea was to include a tree diagram here, why the algorithm came to a specific decision. We discussed this with the experts in the interviews, and we concluded that it might get too fast too complicated to visualize these decisions. Therefore, providing the data sources in this section and statistics seems more beneficial and needs to be implemented in future design iterations.

Success and Failure Rate As mentioned before, there is also a representation of the success and failure rate of the AI's suggestions. This rate depends on the acceptance/ rejection of the AI's impression, done by the user, at the end of the documentation. The AI takes this user input into account under "Verified" or "Rejected" and adapts its certainty score.

Anomaly Detection

Detected anomalies are visualized in the left bar related to their severance. The expert interviews showed that the users do not intuitively understand the visualization of the bars, the colors, and the length of the bars in the original version (see figure 4.12).

Thus in the new design, we used rings to visualize the anomalies; all other (regular) slices are not highlighted in the bar. Depending on this data, the rings are bigger or smaller; this way, the users can see on the first glimpse where to find the critical areas. A visualization utilizing rings seemed more intuitive due to specific areas the user or AI documents in the end.



(a) Version A

(b) Version B

Figure 4.12: Slices-Bar Showing Anomalies Version A and B

New AI Integrated Workflow

To create a fast and intuitive dialogue between user and algorithm, we chose an auto-completion feature. The user starts to write a report, and in that time, the artificial intelligence provides suggestions for auto-completion. The suggestions are also rated with a score by the AI based on experience. The user can immediately accept the first suggestion of the AI (this is also the suggestion the AI is most particular about) or select a finding or parameter from a shortlist of recommendations (also combined with a score of certainty). We chose this workflow because this is the current way radiologists document their findings, and it combines a known strategy with the advancements of artificial intelligence in a familiar way they learn, e.g., from mobile text editors.

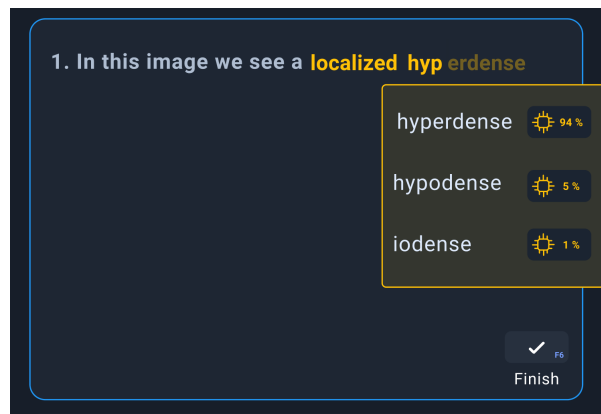


Figure 4.13: Impression, done by the AI

Additionally, the AI gets trained based on the users' input, and this way, the suggestions get already verified or rejected (Fig. 4.10). The highlighting of specific words (or labels) should show the user which keywords are known or essential to the algorithm. After finishing the documentation process, the AI calculates an impression about the finding. This impression can, as mentioned above, be accepted or rejected. When rejected, the impression can be edited and saved. Otherwise, the AI saves the impression, marks the anomaly as verified, and provides information about what it learned through this documentation process.

Visualization of Learning Process

To make transparent to the user what impact the documentation had on the algorithm, it shows what it learned on the way. A list of labels and keywords that the user can still edit at the last moment is shown. This way, the user knows what the AI learned, e.g., "a hyperdense" in the category "finding." These leanings are based on the highlighted keywords/labels in the report text previously written. This visualization should also help the user get feedback from the algorithm and clarify how it works and learns.

4.2 Human in the Loop

The mentioned design adjustments focus on usability, trust, and transparency. Having included the previously mentioned adaptations, cues, and the dialogue between humans and machines, we predict an increase in trust and acceptance towards the usage of the system.

In the new design, it is made more evident to the user the presence of an AI algorithm. We designed the source of the data and its processing more clearly, and at the same time, the new design provides feedback about the quality of the data. We included cues for AI representation in the redesign and implemented an "action route" based on collaboration as suggested by Sundar (2020). The dialogue and the possibility of controlling and manipulating the data should help break up the black box. Additionally, the users should feel more confident being supported by the software. The framework of cues and actions should be a guide to AI trust.

4.3 Formative Evaluation

Before conducting the summative study, a formative evaluation via GoogleMeet with the developing team of deepc helped gather final feedback. One front-end developer, one UX/UI designer, and one of the co-founders of deepc participated in the evaluation meeting. The development of the new prototype was perceived as beneficial and satisfactory to all participants. We visualized the auto-completion feature and the new visualization of anomalies successfully. Although the new features and adaptations would need much time to be developed in the actual product, the participants saw the benefits and improvements that would go along with them. Also, we documented some additional input for future enhancements and partly integrated them already into the prototype after the session, e.g., the feature of showing the rate of true and false predictions done by the machine. After eliminating some minor mistakes, we finalized the prototype for the following summative testing.

Chapter 5

Summative Study

Study Design The study aims to seek answers to the sub questions and address objectives of the thesis:

1. **SQ2:** What makes radiologists trust the output of a structured reporting artificial intelligence solution?
2. **SQ3** To what extent can trust be a key to leverage effective artificial intelligence structured reporting work-flow?
3. **Objective II:** Understand how trust interrelated with artefact acceptance. (trust vs transparency - workflow, visual cues, social cues, content cues)
4. **Objective III:** Provide additional insights on how to design to foster trust and acceptance in Artificial Intelligence.

To answer the research questions and achieve the mentioned objectives, we chose a cross-sectional, between-subject design. The target sample was 12 participants in total. We assigned six participants to group A and the other six to Group B. The study's goal was to measure an in/decrease of usability and trust between A and B. Therefore, we used the independent variable of two prototypes. As dependent variables, we measured: task completion time and task success rate. Additionally, we collected descriptive statistics and scores of the System Usability Scale and the Human-Computer Trust Scale.

5.1 Instruments and Methods

Method	Research Question (RQ) / Sub-question (SQ) / Objectives (OB)	Comments
<p>Tasks Prototype A or B Group A functions as control group, independent variable: Design variation</p>	-	Group A with experiment A followed with measured observation Group B with experiment B followed with measured observation
<p>Human Computer Trust Scale (HCTS) dependent variable: perception of trust</p>	<p>SQ2 What makes radiologists trust the output of a structured reporting artificial intelligence solution?</p> <p>SQ3 To what extent can trust be a key to leverage effective artificial intelligence structured reporting work-flow?</p>	
<p>System Usability Scale (SUS) dependent variable: perception of usability + time of completion</p>	<p>SQ1 How can we design a structured reporting artificial intelligence solution to support the clinical work-flow of radiologists?</p>	

Retrospective Interview structured	SQ1 How can we design a structured reporting artificial intelligence solution to support the clinical work-flow of radiologists? SQ2 What makes radiologists trust the output of a structured reporting artificial intelligence solution?	The possibility to qualitatively dive deeper.
--	--	---

Prototype

1. Prototype A (current version of deepcOS)
2. Prototype B (redesigned version of deepcOS with, adaptations regarding usability, inclusion of trust cues and more transparency)

Survey

Similar to the surveys in the expert's interviews, an online survey including the Human Computer Trust Scale (HCTS) developed by Gulati et al. (2019) and the System Usability Scale (SUS) (?) was presented to the users after experiencing the prototype. To gather this data we used, Loop11 as a tool.

Structured Interview

We included two additional questions after the SUS and HCTS survey. To determine which of the new features related to AI were accepted the most, one question was added to ask the users to rate their most exciting feature. There the participants should select their preferences. Since there was no relevant AI representation in Version A, we asked this question only in study B. Additionally, the asked all participants for further recommendations and comments.

Demographics

To collect some demographics, the following parameters we chose: age, gender, occupation / in years, institution, assessment in the technical experience. Interesting for the study would be if differences can be observed regarding the institution. Most of the expertise relied on experts from the university clinic environment. Therefore it would be interesting if there are differences between private intuitions or state-driven hospitals.

5.2 Participants

To recruit participants for the study, we contacted several radiologists and institutions via e-mail or LinkedIn. Therefore a "Call for Study" paper was written in German and English to be sent out to possible participants. The document can be viewed in the appendix section. It includes essential background information about the cause of the study and details about the study itself. Due to the ongoing Covid19 pandemic and the small population, it seemed challenging to recruit radiologists and even medical staff. Therefore the sample population was reduced to 12 participants and extended to experts who develop diagnostic tools driven by AI.

One medical student, one physiotherapist, one occupational therapist participated in the study, and nine participants worked in the health industry. The participants working in the health industry can be divided into product owner (1), communication specialist (3), UX designer (1), medical device vendor (1), and AI developers (3).

5.3 Study

5.3.1 Risks to Validity

During the development of the study, several issues occurred that might risk the validity of the study. The loop11 software provides high functionality regarding analyzing and reporting. However, in guiding the users through the survey, there are multiple issues, e.g., the users need to open up the task description manually to click on a "task completed" button instead of being guided automatically to

the next task. This was solved by including an additional task to start with. This task should teach the participants how to walk through the study. In general, unmoderated remote testing could be a risk since the participants have no moderator to guide them if there are technical problems or if they get lost in the prototype. All these factors need to be taken into account. Nevertheless, all participants' screens were recorded during the session for documentation to be used as a backup.

5.3.2 Pilot Test

The pilot test was conducted with five participants. These participants tested the study process based on technical feasibility, understanding, flawless walk-through, and spelling. We included the feedback provided by the pilot testers in the study. Besides spelling and minor mistakes, one improvement to the study was defining the task description. Since the users walk through unmoderated remote testing and due to the limitations of the testing software, it is mandatory to have clear explanations.

5.4 Findings

5.4.1 Task Success and Time on Task

A clear starting and ending point was defined and communicated to the user to measure task success and time. Since the participants were facing an unmoderated session, we needed to express the instructions of starting and ending points, e.g., *"The task can be seen as completed when you see a brain scan with blue highlighted areas."*(see figure 5.1). The testing application used to guide the users through the testing provides automated tracking of the task success and time on task parameters. During the study design, we needed to set a start and end screen in the loop11 software. Then the parameters were collected after entering this start screen until the task was marked as completed by the user.

Overall

The overall task success rate for the tasks of prototype A was 33 % success and 20 % fail (see figure 5.2). For prototype B, we observed an increase in the success rate to 53% and an increase of failed

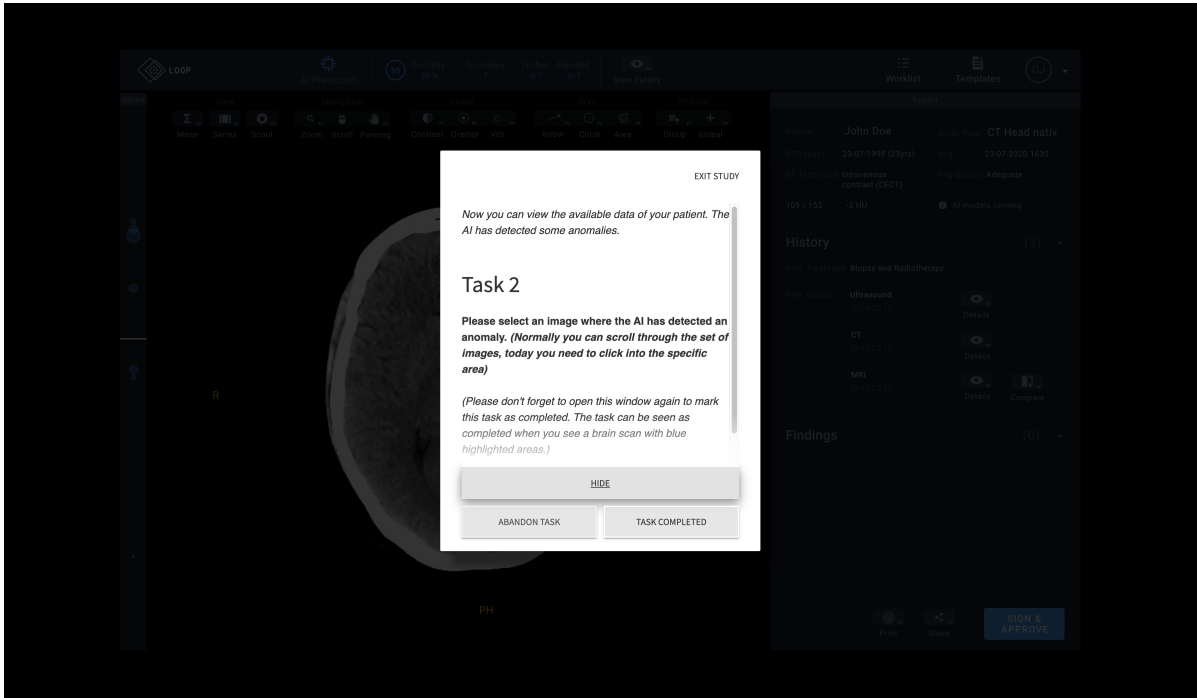


Figure 5.1: Introductions to the "Measurements" of prototype B

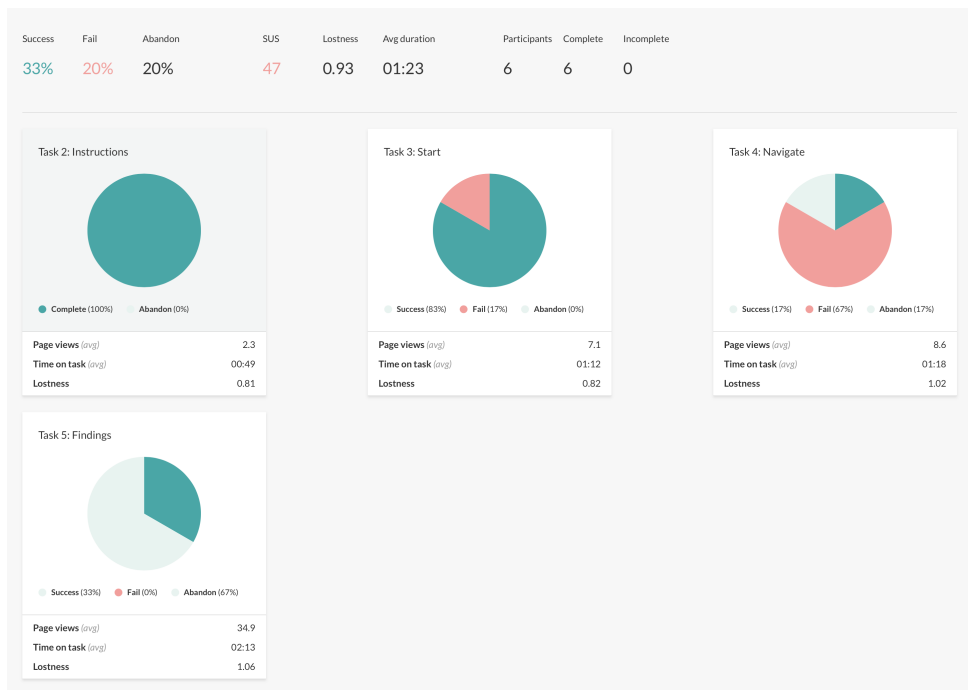


Figure 5.2: Results testing version prototype A

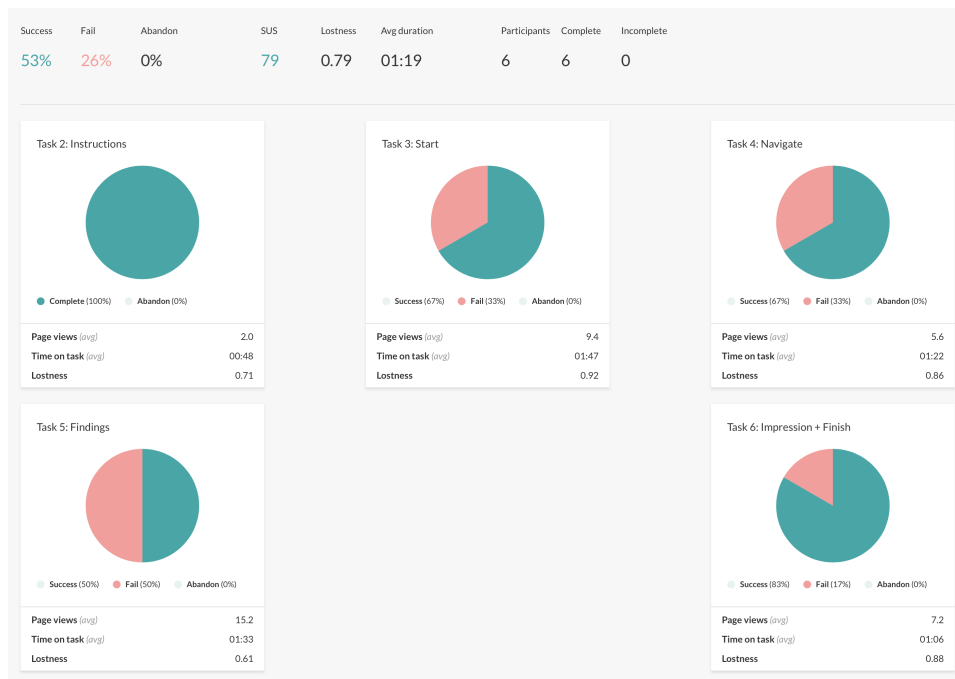


Figure 5.3: Results testing version prototype B

tasks to 26 % (see figure 5.2). We measured the average duration on a task of prototype A at 01:23 minutes which was slightly higher than the average time spent on tasks of prototype B, which was 01:19 min. We can see a clear improvement in abandoned tasks; the participants abandoned 20 % of the tasks during the study using prototype A. None of the tasks were abandoned using prototype B. All twelve participants finished the introductory task with success.

Task "Start"

In their first tasks, the participants should log into the system using provided credentials. Furthermore, they should select a patient from a list that has not been reviewed yet. 83 % of the participants were successful, and 17% failed in finishing this task using prototype A. Prototype B did not provide an improvement here; a decrease of success and an increase of failure, each by 16%, was documented. Also, the average duration spent on this task increased by 00:35 min from 01:12 for prototype A and 01:47 min for prototype B.

Task "Navigate"

During the second task, the participants were asked to select an image where the AI has detected an anomaly. 17 % of the participants were successful, and 67% failed in finishing this task using

prototype A. 17 % abandoned this task during the process. In this task, we see a clear improvement in the success rate, using prototype B, by 50% to 67%. 33% failed the task, and none of the users of prototype B abandoned the task. Nevertheless, the time spent on the task increased slightly in prototype B.

The average duration spent on prototype A was 01:18 min, while the time increased in prototype B by 00:04 min to 01:22 min.

Task "Findings"

In the following task, the user was asked to draw an arrow in their region of interest, write their documentation until they had all information in their report section and finally finish the documentation process. In prototype A, this task was successfully finished by 33% of the participants, while 0% failed in completing this task. 67% abandoned the task during the process. These tasks also showed improvements for prototype B; the task was successfully finished by 50%, which is an increase of 17%. 50% still failed to complete the task, but none of the users abandoned the process. The average duration spent on this task improved significantly in prototype B, measured with 01:33 min, an improvement of 00:40 min to prototype A, measuring 02:13 min for this task.

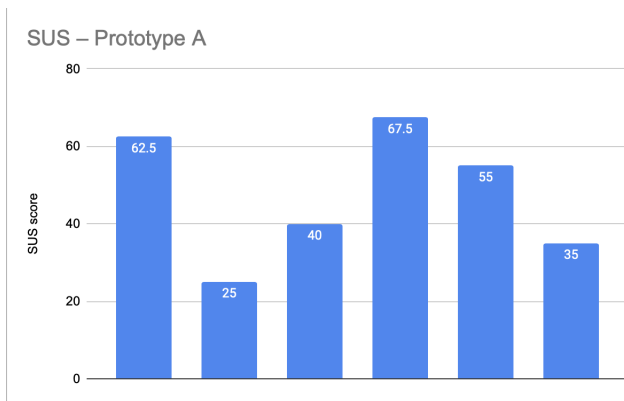
Task "Impression + Finish"

We included an additional task in the prototype to increase transparency in the workflow, which was only available in prototype B. In their last tasks, the participants should accept or reject the AIs impression and sign/approve the report for this finding. This task was successfully finished by 83 % of the participants, while 17% failed in completing this task. The average duration spent on this task was 01:06 min.

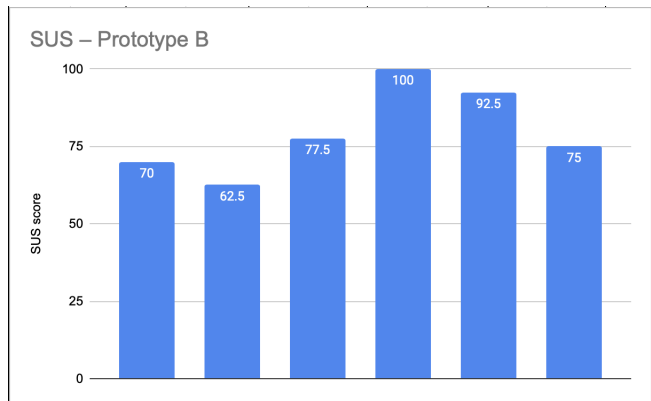
5.4.2 Survey

System Usability Scale

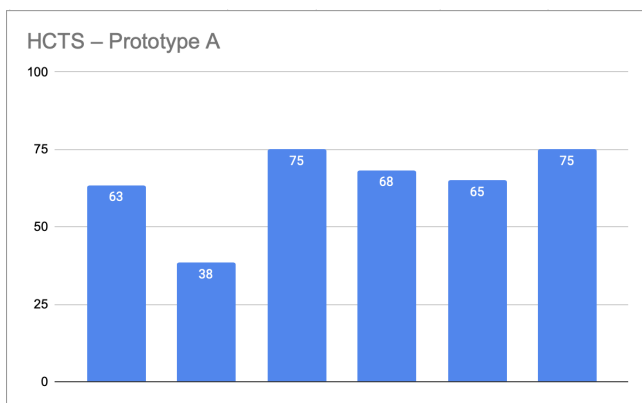
The average score of the SUS method between all the six participants of prototype A is 47. This result is considered below-the-average within the industry. The average score is 68. The lowest rating by



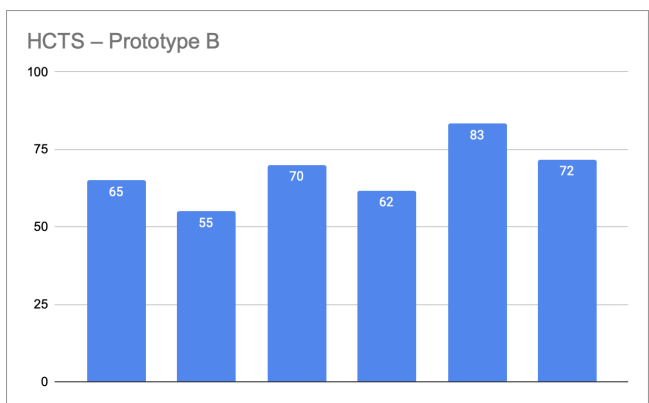
(a) SUS ratings prototype A



(b) SUS ratings prototype B



(a) HCTS evaluation prototype A



(b) HCTS evaluation prototype B

one of the participants was 25 and the highest 67,5, which still is below average. The average score for prototype B has been rated to 79. We see this result as a clear improvement in the perception of usability. The lowest rating given by the participants was 62.5 and the highest 100. The following diagram visualizes the results from all the 12 participants for prototype A and prototype B.

Human Computer Trust Scale

Evaluating the Human Computer Trust Scale for the two prototypes resulted in a slightly higher trust score, by 4%, in prototype B. Prototype A received an average rating of 64% and prototype B received 68%, which is below average (which would be 75%). The detailed data (view following tables) visualize that in prototype B the assessment in reciprocity decreased about 1% (from 76% for prototype A and 75% for prototype B). Benevolence increased about 12% to 92% for prototype B (from 80% for prototype A), which we see as a significant improvement. We could observe the most considerable progress in the assessment of competence. Due to the ratings, the evaluation of competence increased about 20% for prototype B to 90% (from 70% for prototype A).

The participants lowered their general assessment of trust of prototype B from 79% (prototype A) to 65% of about 14%.

HCTS - Prototype A	Average	Standard Deviation	Trust Score
Reciprocity	3.2	0.7231847957	76%
Benevolence	3.3	0.8740760794	80%
Competence	3.0	1.019759303	73%
General Trust Measurement	3.3	0.8720617474	79%

HCTS - Prototype B	Average	Standard Deviation	Trust Score
Reciprocity	3.1	1.25347229	75%
Benevolence	3.8	0.7316338187	92%
Competence	3.9	0.8174358796	93%
General Trust Measurement	2.7	0.8491945995	62%

5.4.3 Qualitative Questions

The participants evaluating prototype B were asked some additional questions regarding AI integration. The results should help identify the most promising features and cues communicating AI processes to the users.

When asked: "What was the most promising feature regarding AI in your eyes?" four of the six participants selected "The AI provides text completion" and "The AI shows me anomalies." Three participants selected "The AI provides an impression" and "The AI provides information about its certainty." One user liked the feature "In the End, the AI showed what it learned" as a promising feature (see figure 5.6).

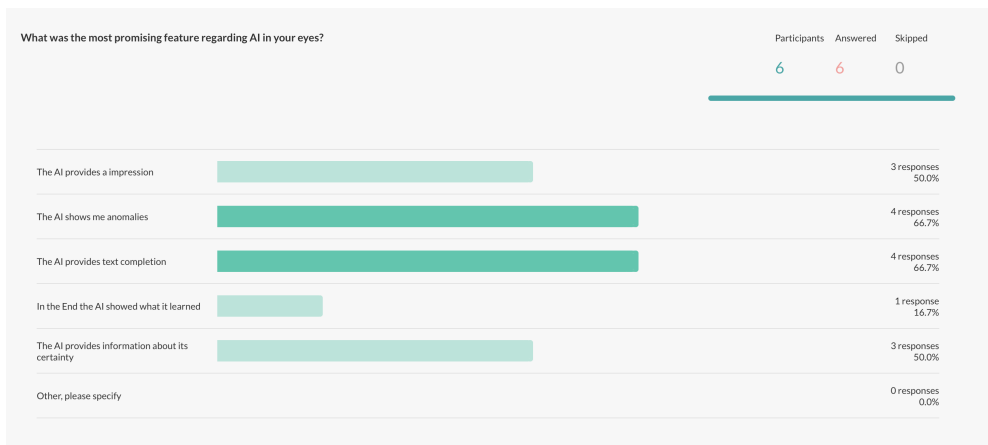


Figure 5.6: Results of "What was the most promising feature regarding AI in your eyes?"

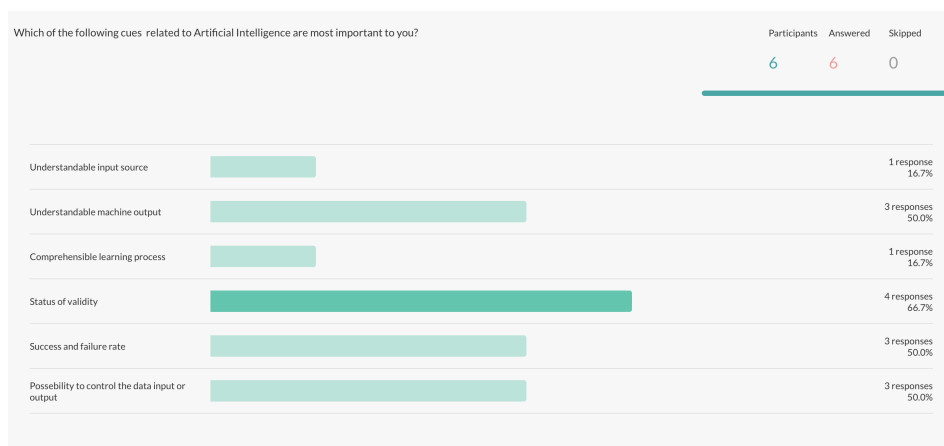


Figure 5.7: Results of "Which of the following cues related to Artificial Intelligence are most important to you?"

Regarding "Which of the following cues related to Artificial Intelligence are most important to you?" Four of the six participants selected "Status of validity" as the most promising cue. Three users set "Success and failure rate," "Possibility to control the data input or output," as well as "Understandable machine output." Only one participant, each, selected "Comprehensible learning process" or "Understandable input source" as necessary (see figure 5.7).

5.5 Design Recommendations

Research Question *How can machine learning processes be designed to foster trust in users?*

Since all the sub-questions of the study have been addressed, the following section will describe the design process's rationale and design recommendations (see figure 5.8). These recommendations should support practitioners in the medical domain as well as in the HCI community in the task of designing for AI-driven software.

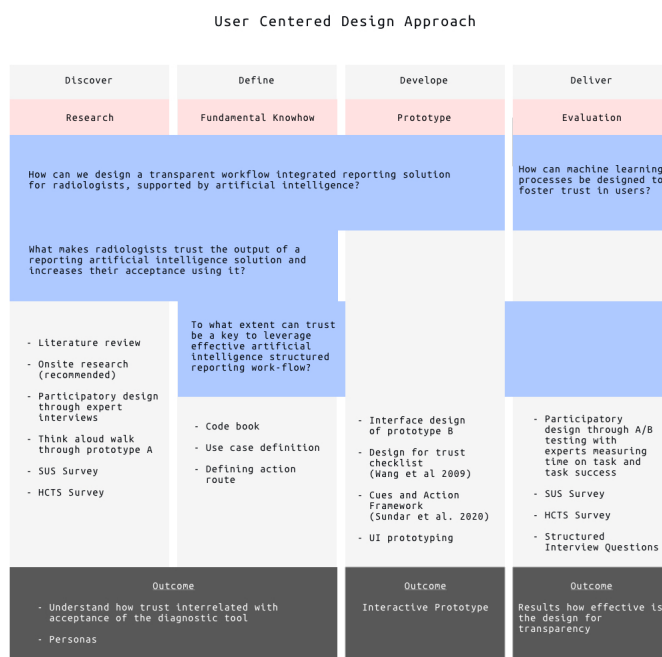


Figure 5.8: Final Design Process

In the discover phase the participatory user-centered design approach started with theoretical research and integrating actual users in the early process in the discovery and definition phase of the study. The information of the experts was essential to focus on the most critical features and procedures in a radiologist's daily workflow. The interviews were used to determine what trust in AI means to the users and what would be needed to enhance their trust in AI-driven image diagnostic tools. In the developing phase, a new design was created, building upon the gathered data. This way, we increased our understanding of how trust interrelates with acceptance of the diagnostic tool.

The experts' perception of trust regarding AI and what could make them trust more in AI-driven systems helped to define the cues used in the design. All of the experts mentioned the increase of transparency as the main factor to enhance trust. Furthermore, visualize where the data comes from or provide certainty of the model output. The assessment of the status quo of deepcOS helped to find potential improvements in usability as well. In the summative study, we observed that the cues integrated into the new design improved trust regarding benevolence and competence. Reciprocity and the overall trust in the system still need improvement. Therefore the signals with the most impact regarding study participants were status of validity, success and failure rate, and the possibility to control the data. All three cues are related to the quality of the model output. Cues related to the input source and learning status were not equally relevant to the users. The findings match the experts' opinions, but additional information regarding the input source would be necessary (e.g., preliminary studies, the validity of the developing company).

The summative study let us measure how effective the design for transparency changed the assessment of trust and usability of the diagnostic tool (OBJ I). This helped us also to answer sub-question 3.

Proposal for enhanced trust design
checklist for AI enhanced software

Dimension	Checklist Item	Positive	Negative
VISUAL	Professional design		
	Color scheme to suit the product/service		
	Nice and legible fonts		
	High-quality (and authentic) images and visuals		
	Good on-site search		
	Easy-to-use navigation		
	Clear anchor text and microcopy		
CONTENT DESIGN	No technical problems (broken links, missing pictures or pages, etc.)		
	Brand-promoting information (Logo, slogan)		
	Company information ("About" page, facts & figures)		
	Contact information		
	Useful (expert-level) content		
	Good grammar and minimized jargon		
	External links (sources)		
	Up-to-date blog		
	Changelog		
	Developer case studies		
	Clear Terms of Service (Developer oriented that differ from company's own ToS)		
	Helpful FAQs		
	Provide clean and working code examples		
DEVELOPER CONTENT DESIGN	Provide relevant background knowledge		
	Connect concepts to code		
	Enable selective access to code		
	Signal text-to-code connections.		
	Provide important information redundantly		
	Enable fast use of the API		
SOCIAL-CUE DESIGN	Easy access to customer service (e.g., contact form)		
	Developer Testimonials from successful businesses that built an integration		
	Reviews from influencers and notable developers		
	Social presence (Developer community, social media)		
AI-CUE DESIGN	AI presence if critical to the product		
	Understandable input source		
	Understandable machine output		
	Comprehensible learning process		
	Status of validity		
	Success and failure rate		
	Possibility to control the data input or output		

Figure 5.9: Adapted Design Trust Checklist by Wang (2009) enhanced with AI suggestions

As mentioned before, we did not change the general design language itself in this new prototype. But we made several adaptations to improve the overall usability and include cues that should increase transparency and foster trust in the diagnostic tool. During the process, we followed the framework of cues and actions by Sundar (2020) and an adapted version of the design for trust checklist based on the developments of Wang (2009). The dialogue and the possibility of controlling and manipulating the data should help break up the black box. Additionally, the users should feel more confident being supported by the software. Based on the summative testing results, the author recommends enhancing the design for the trust checklist with a section for AI integration (see figure 5.9).

Chapter 6

Discussion

In the discovery phase, the initial qualitative interviews revealed that the design and the usability of deepcOS are already on a satisfying level. The triangulation of data collection can confirm this through the positive results of the SUS questionnaire (68,4 score), the outcome of the semi-structured interview, and the average expert rating of trust of 74%. The expert interviews provided valuable input and helped identify risks, opportunities, features, needs, and expectations from real users. Conducting the interviews with the experts seemed to be the most valuable input in this research. This confirms the importance of speaking with real users. Therefore the expert interviews are seen as an essential input in this research. The new prototype focused on the visualization of AI processes and the communication of AI output. We integrated features and cues that provided loops between humans and the machine and enhanced trust in the system. Additionally, some changes regarding layout and workflow improved the overall usability of the tool. Constructive discussions with the developing team of deepc provided critical feedback and helped to integrate additional features and eliminate minor mistakes in the prototype. The support of the experts and the team of deepc was beneficial in providing valuable domain knowledge in the field.

Reviewing the summative testing results, we observed an improvement in task success and time on task for the new design. The task success rate improved by about 20% and the time on task improved by four seconds. These results conclude that the new design lets the user achieve their goals quicker and easier. The improved SUS score supports this assumption. According to the SUS results, the usability improved about 32 (49 for prototype A and 79 for prototype B). The HCTS score only improved slightly in the new design, of about 4% (64% for prototype A and 66% for prototype B),

and remains below average. When reviewing the results for trust in prototype A, we can see a big difference in the experts' results. In the study, prototype A was given lower rates by the participants than rated by the experts. This deviation might have been caused by the change of the target user group for the study from radiologists to AI specialists and other potential users from the medical domain. Another explanation could be that the experts walked through a moderated testing and the participants of the summative study faced an unmoderated session. This might have led to lower results in the SUS and HCTS Scores. It leads to the conclusion, like apprehended in chapter "Risks to Validity," that the study needs to be conducted in a follow-up study with the target user group, radiologists, to get more authentic results. Thus the summative study can be seen as a first validation of the new design suggestions and lead to the first design recommendations. Using the trust cues and new features in the new design helped to improve the usability significantly and partly the perception of trust in the diagnostic tool. Therefore we can say that increasing transparency helps to increase effectiveness and even efficiency in the workflow. This statement is strengthened by the rise in the SUS score and the task success and time on task results that improved in the new design. Due to the experts, the increase in transparency should also increase trust in the system. This was only particularly the case for the new design. But we can say that the increase of transparency in AI systems and processes increases effectiveness and efficiency in the AI structured reporting workflow.

Limitations and Opportunities for Future Research The author is aware that there are more possibilities to design for transparency when it comes to AI integration. Additionally, due to the ongoing Covid19 pandemic, we could not conduct sufficient onsite user research. Although expert interviews lead to first insights, not having visited and observed real users in their working environment leads to limitations in understanding and overlooking the whole user workflow in which the AI and machine learning processes should be integrated. Thus the author recommends visiting onsite radiologists in different locations in the future to get diverse and authentic insights. Due to the extremely small sample size of high specialized users and the ongoing circumstances, it was tough to recruit target users. Therefore it is strongly recommended to conduct a follow-up study with several radiologists to validate the new design and get more authentic results. The application loop¹¹ had limitations, too. The users needed to confirm their tasks as finished upon their perception. This risk of validity was handled by providing exact information about the state of completion. The application had some minor bugs during the process as well. Thus, to prevent errors caused by the testing application, it is recommended by the author to conduct a follow-up study using moderated sessions.

Chapter 7

Conclusion

This case study aimed to provide design recommendations to enhance usability and trust in using a medical image diagnostic tool supported by artificial intelligence. We wanted to find out how can machine learning processes be designed to foster trust in users? The participatory design approach helped to develop solutions based on real users' needs. Thus we can say that the research provided valuable insights into the opportunities and barriers of AI-driven workflows. Based on these insights, we developed a new design that leads to the first conclusions and recommendations regarding AI cues and features that improve usability and foster trust in AI-supported tools. This case study leads to the first assumption that we successfully integrated loops between humans and machines to create a dialogue for input and output. Focusing on designing for transparency, we managed to increase the usability and trust related to benevolence and competence. The development of medical diagnostic tools and other software created by the HCI community can benefit from the process described in this research and the adapted Design Trust Checklist. The provided framework leads to a transparent design that seeks to avoid the "black box" problem, creating information loops that keep the human in the loop.

Chapter 8

References

References

- 5 key takeaways from a patient survey about AI in radiology. (n.d.). Retrieved from <https://www.radiologybusiness.com/topics/artificial-intelligence/5-key-takeaways-survey-ai-radiology>
- Bidgood, W. D. (1998). Clinical importance of the DICOM structured reporting standard. *International Journal of Cardiac Imaging*, 14(5), 307–315. Retrieved from <https://pubmed.ncbi.nlm.nih.gov/10453383/> doi: 10.1023/a:1006073709957
- Chartrand, G., Cheng, P. M., Vorontsov, E., Drozdal, M., Turcotte, S., Pal, C. J., ... Tang, A. (2017). Deep Learning: A Primer for Radiologists. <https://doi.org/10.1148/rg.2017170077>, 37(7), 2113–2131. Retrieved from <http://pubs.rsna.org/doi/10.1148/rg.2017170077> doi: 10.1148/RG.2017170077
- Chen, T. W., & Sundar, S. S. (2018, 4). "This app would like to use your current location to better serve you": Importance of user assent and system transparency in personalized mobile services. In *Conference on human factors in computing systems - proceedings* (Vol. 2018-April). Association for Computing Machinery. doi: 10.1145/3173574.3174111
- Dikici, E., Bigelow, M., Prevedello, L. M., White, R. D., & Erdal, B. S. (2020, 2). Integrating AI into radiology workflow: levels of research, production, and feedback maturity. *Journal of Medical Imaging*, 7(01), 1. Retrieved from <https://www.spiedigitallibrary.org/terms-of-use> doi: 10.1117/1.jmi.7.1.016502
- Einführung, E., & ö Springer, A. V. (n.d.). *Michael Häder Empirische Sozialforschung* (Tech. Rep.).
- Fimberg, K., & Sousa, S. (2020, 10). The Impact of Website Design on Users' Trust. In *Acm international conference proceeding series* (pp. 1–5). New York, NY, USA: Association for Computing Machinery. Retrieved from <https://dl.acm.org/doi/10.1145/3419249.3420086> doi: 10.1145/3419249.3420086
- Fukushima, K., & Miyake, S. (1982). Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Visual Pattern Recognition. In (pp. 267–285). Springer, Berlin, Hei-

delberg. Retrieved from https://link.springer.com/chapter/10.1007/978-3-642-46466-9_18 doi: 10.1007/978-3-642-46466-9{_}18

Gueld, M. O., Kohnen, M., Keyzers, D., Schubert, H., Wein, B. B., Bredno, J., & Lehmann, T. M. (2002, 5). *Quality of DICOM header information for image categorization*. In *Medical imaging 2002: Pacs and integrated medical information systems: Design and evaluation* (Vol. 4685, pp. 280–287). SPIE. doi: 10.1117/12.467017

Gulati, S., Sousa, S., & Lamas, D. (2018, 12). Modelling trust in human-like technologies. In *Acm international conference proceeding series* (pp. 1–10). New York, NY, USA: Association for Computing Machinery. Retrieved from <https://dl.acm.org/doi/10.1145/3297121.3297124> doi: 10.1145/3297121.3297124

Gulati, S., Sousa, S., & Lamas, D. (2019, 10). Design, development and evaluation of a human-computer trust scale. *Behaviour & Information Technology*, 38(10), 1004–1015. Retrieved from <https://www.tandfonline.com/doi/full/10.1080/0144929X.2019.1656779> doi: 10.1080/0144929X.2019.1656779

Harvey, H., & Glocker, B. (2019, 1). A standardised approach for preparing imaging data for machine learning tasks in radiology. In *Artificial intelligence in medical imaging: Opportunities, applications and risks* (pp. 61–72). Springer International Publishing. Retrieved from https://link.springer.com/chapter/10.1007/978-3-319-94878-2_6 doi: 10.1007/978-3-319-94878-2{_}6

Impact of artificial intelligence on radiology: a EuroAIM survey among members of the European Society of Radiology. (2019). *Insights into Imaging 2019 10:1*, 10(1), 1–11. Retrieved from <https://insightsimaging.springeropen.com/articles/10.1186/s13244-019-0798-3> doi: 10.1186/S13244-019-0798-3

Jhaver, S., Karpfen, Y., & Antin, J. (2018, 4). Algorithmic anxiety and coping strategies of airbnb hosts. In *Conference on human factors in computing systems - proceedings* (Vol. 2018-April). Association for Computing Machinery. doi: 10.1145/3173574.3173995

Kim, K. J. (2016, 11). Interacting Socially with the Internet of Things (IoT): Effects of Source Attribution and Specialization in Human–IoT Interaction. *Journal of Computer-Mediated Communication*, 21(6), 420–435. doi: 10.1111/jcc4.12177

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data and Society*, 5(1). doi: 10.1177/2053951718756684

- Levac, D., Colquhoun, H., & O'Brien, K. K. (2010, 9). Scoping studies: Advancing the methodology. *Implementation Science*, 5(1). doi: 10.1186/1748-5908-5-69
- Liao, Q. V., Gruen, D., & Miller, S. (2020, 1). Questioning the AI: Informing Design Practices for Explainable AI User Experiences. *Conference on Human Factors in Computing Systems - Proceedings*. Retrieved from <http://arxiv.org/abs/2001.02478><http://dx.doi.org/10.1145/3313831.3376590> doi: 10.1145/3313831.3376590
- Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., ... Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42, 60–88. Retrieved from <https://linkinghub.elsevier.com/retrieve/pii/S1361841517301135> doi: 10.1016/J.MEDIA.2017.07.005
- Lo, S. C. B., Lou, S. L. A., Chien, M. V., & Mun, S. K. (1995). Artificial Convolution Neural Network Techniques and Applications for Lung Nodule Detection. *IEEE Transactions on Medical Imaging*, 14(4), 711–718. doi: 10.1109/42.476112
- Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995, 7). An Integrative Model of Organizational Trust. *The Academy of Management Review*, 20(3), 709. doi: 10.2307/258792
- Mcknight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011, 6). Trust in a specific technology. *ACM Transactions on Management Information Systems*, 2(2), 1–25. Retrieved from <https://dl.acm.org/doi/10.1145/1985347.1985353> doi: 10.1145/1985347.1985353
- Med Students and AI June 2019* — American College of Radiology. (n.d.). Retrieved from <https://www.acrdsi.org/Blog/Med-Students-and-AI-June-2019>
- Morville, P. (n.d.). *User Experience Design*. Retrieved from http://semanticstudios.com/user_experience_design/
- O., A., & A., W. (n.d.). *Towards algorithmic experience: Initial efforts for social media contexts. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI'18)*.
- Olthof, A. W., Leusveld, A. L., de Groot, J. C., Callenbach, P. M., & van Ooijen, P. M. (2020, 9). Contextual Structured Reporting in Radiology: Implementation and Long-Term Evaluation in Improving the Communication of Critical Findings. *Journal of Medical Systems*, 44(9). Retrieved from <https://pubmed.ncbi.nlm.nih.gov/32725421/> doi: 10.1007/s10916-020-01609-3
- Rader, E., Cotter, K., & Cho, J. (2018, 4). Explanations as mechanisms for supporting algorithmic transparency. In *Conference on human factors in computing systems - proceedings* (Vol. 2018-

- April). Association for Computing Machinery. doi: 10.1145/3173574.3173677
- Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). *Not so different after all: A cross-discipline view of trust* (Vol. 23) (No. 3). doi: 10.5465/AMR.1998.926617
- Sendak, M., Elish, M., Gao, M., Futoma, J., Ratliff, W., Nichols, M., ... O'Brien, C. (2019). "The Human Body is a Black Box": Supporting Clinical Decision-Making with Deep Learning. Retrieved from <http://arxiv.org/abs/1911.08089>
- Strohm, L., Hehakaya, C., Ranschaert, E. R., Boon, W. P., & Moors, E. H. (2020, 10). Implementation of artificial intelligence (AI) applications in radiology: hindering and facilitating factors. *European Radiology*, 30(10), 5525–5532. Retrieved from <https://doi.org/10.1007/s00330-020-06946-y> doi: 10.1007/s00330-020-06946-y
- Sundar, S. S. (2020, 3). Rise of Machine Agency: A Framework for Studying the Psychology of Human–AI Interaction (HAI). *Journal of Computer-Mediated Communication*, 25(1), 74–88. Retrieved from <https://academic.oup.com/jcmc/article/25/1/74/5700811> doi: 10.1093/jcmc/zmz026
- Wang, Y. D. (2009). *Trust in B2C E-Commerce Interface* (Tech. Rep.). Retrieved from <https://www.researchgate.net/publication/243443945>
- Williams, L., Carrigan, A., Auffermann, W., Mills, M., Rich, A., Elmore, J., & Drew, T. (2020). The invisible breast cancer: Experience does not protect against inattentive blindness to clinically relevant findings in radiology. *Psychonomic Bulletin and Review*. doi: 10.3758/s13423-020-01826-4
- Yu, K. H., & Kohane, I. S. (2019, 3). *Framing the challenges of artificial intelligence in medicine* (Vol. 28) (No. 3). BMJ Publishing Group. Retrieved from <https://qualitysafety.bmj.com/lookup/doi/10.1136/bmjqs-2018-008551> doi: 10.1136/bmjqs-2018-008551
- Zhang, B., & Sundar, S. S. (2019, 8). Proactive vs. reactive personalization: Can customization of privacy enhance user experience? *International Journal of Human Computer Studies*, 128, 86–99. doi: 10.1016/j.ijhcs.2019.03.002

Chapter 9

Appendix

9.1 User Study Plot: Expert Interviews

1 / USER STUDY PLOT – RADIOLOGIST



① WELCOME AND INTRODUCTION (5 min)

- + Warm up (a few simple questions to ease tension)
- + Aim of the study: Development of a software that evaluates radiological imaging using Artificial Intelligence.
- + Methodology and documentation consent: Explain research methods to radiologist, get permission to take pictures, video snippets and eye-tracking (if a prototype is shown), the study is recorded for documentation (anonymous)

② CONSENT FORM (2 min)

- + Signed between the respondent and deepc representatives

③ deepcOS PROTOTYPE (10 min)

- + Self Exploration
- + Explanation and answering questions

④ INTERVIEW I + II (+DEMOGRAPHIC) (Semi structured interview in a quite space) (30 min) +

- + Questionnaire: Human Computer Trust Scale (HCTS)
- + Questionnaire: System Usability Scale (SUS)

- + General questions: Personal background, work-flow
- + Structured Reporting: personal work-flow, barriers
- + Artificial Intelligence: experience, assessment
- + deepcOS: assessment, opinion (if a prototype is shown)
- + General questions: working environment, personal preferences (cool down questions)
- + Demographical questionnaire: personal information

⑤ THANK YOU & GOODBYE

- + Debriefing session (Team): Noting down five key findings

2 DECLARATION OF CONSENT



RESEARCH DECLARATION OF CONSENT

Study: _____

Name: _____

PARTICIPANT

I was sufficiently informed orally for me and / or written of the interview with deepc.
I hereby agree that during the study informations will be recorded (audio and video).
This information is used solely to improve our products / prototypes. I understand that my participation is kept confidential in this study, and I am even obliged to maintain confidentiality regarding study content, product information and the used product prototypes.
All personal information and individual results will not be disclosed to third parties.

deepc

No information regarding patients (foto, video and data) will be published and will be disguised.

I understand that I may withdraw at any time from participation in the study.

With the described procedure I agree and confirm this with my signature.

Date: _____

Signature: _____

Date: _____

Signature: _____

3 QUESTIONNAIRE (HCTS)



YOUR OPINION

1 2 3 4 5

I believe that there could be negative consequences when using deepcOS

Strongly disagree Strongly agree

1 2 3 4 5

I feel I must be cautious when using deepcOS

Strongly disagree Strongly agree

1 2 3 4 5

It is risky to interact with deepcOS

Strongly disagree Strongly agree

1 2 3 4 5

I believe that deepcOS will act in my best interest

Strongly disagree Strongly agree

1 2 3 4 5

I believe that deepcOS will do its best to help me if I need help

Strongly disagree Strongly agree

1 2 3 4 5

I believe that deepcOS is interested in understanding my needs and preferences

Strongly disagree Strongly agree

3 QUESTIONNAIRE (HCTS)



YOUR OPINION

1 2 3 4 5

I think that deepcOS is competent and effective in image interpretation

Strongly disagree Strongly agree

1 2 3 4 5

I think that deepcOS performs its role as supporting software very well

Strongly disagree Strongly agree

1 2 3 4 5

I believe that deepcOS has all the functionalities I would expect from a software that should support image interpretation

Strongly disagree Strongly agree

1 2 3 4 5

If I use deepcOS, I think i would be able to depend on it completely

Strongly disagree Strongly agree

1 2 3 4 5

I can always rely on deepcOS for image interpretation

Strongly disagree Strongly agree

1 2 3 4 5

I can trust the information presented to me by deepcOS

Strongly disagree Strongly agree

3 QUESTIONNAIRE (SUS)



YOUR OPINION

1 2 3 4 5

I think that I would like to use this system frequently.

Strongly disagree Strongly agree

1 2 3 4 5

I found the system unnecessarily complex.

Strongly disagree Strongly agree

1 2 3 4 5

I think that I would need the support of a technical person to be able to use this system.

Strongly disagree Strongly agree

1 2 3 4 5

I found the various functions in this system were well integrated.

Strongly disagree Strongly agree

1 2 3 4 5

I thought there was too much inconsistency in this system.

Strongly disagree Strongly agree

1 2 3 4 5

I would imagine that most people would learn to use this system very quickly.

Strongly disagree Strongly agree

3 QUESTIONNAIRE (SUS)



YOUR OPINION

1 2 3 4 5

I found the system very cumbersome to use.

Strongly disagree Strongly agree

1 2 3 4 5

I felt very confident using the system.

Strongly disagree Strongly agree

1 2 3 4 5

I needed to learn a lot of things before I could get going with this system.

Strongly disagree Strongly agree

4 INTERVIEW



Please record the interview!

Thank you for your patience.

GENERAL

1. Please describe your personal work-flow

(First steps, check patient, read image, documentation)

2. How much time do you have for one patient (estimation).

3. Please describe the modalities (CT, MRI, etc.) and the body parts (Brain, Chest, Abdomen) you are diagnosing most commonly. (Work-flow with medical assistant)

4 INTERVIEW



WORK-FLOW

4. Which of your tasks in your work-flow is the most time consuming and why?

(Reading image, documentation)

5. How do you think could this task be simplified?

(Ideas about possible solutions)

6. Please describe the biggest barriers you encounter in your work-flow.

(Slow software, inconsistencies, user unfriendly documentation...)

4 INTERVIEW



STRUCTURED REPORTING

7. Please describe your current workflow of reporting.

8. In which form are you entering your report? (Dictating, Writing)

9. Are you familiar with structured reporting? (If not, needs to be explained)

10. What do you think about this kind of reporting?

4 INTERVIEW



11. In what way do you think modular reporting can improve the status quo? (If the user does not have answers, ask more concretely: Completeness? Time efficiency? Standardization? (between colleagues) Connection between image and finding?)

12. What do you think are the drawbacks?

ARTIFICIAL INTELLIGENCE

13. Where have you utilised AI-based products or services in your clinical practice?

14. Where do you think could AI-based products or services support you in your clinical practice?

4 INTERVIEW



15. What would be the role of radiologists in developing / validation AI applications to medical imaging?

16. Would you invest time to improve AI-algorithms e.g. via providing feedback?

17. What would be necessary for you to trust the outcome of the AI-prediction?

4 INTERVIEW



18. How do you think would controlling the data or training process of the machine affect your trust perception regarding an AI-service?

deepcOS

19. Please describe your first impression with the software.

20. What did you like about the software?

21. How do you think the software can support you in your individual needs?

4 INTERVIEW



22. What seemed odd to you or did you dislike going through the software?

23. What features did you miss that are not implemented in deepcOS?

General

24. In which moment, during your work did you reach your limits?

25. What did you miss in that specific moment?

4 INTERVIEW



26. How does the environment affect your work?

27. What could be improved from your perspective to make your personal workflow more simple and efficient?

4 INTERVIEW PART II (DEMOGRAPHY)



PERSONAL DETAILS

1 2 3 4 5

Own assessment of technical experience (with electronic products):

not experienced at all very experienced

Occupation / Years:

Age:

Gender:

Highest Reached
Qualification:

**Additional
Comments**

9.2 User Study Plot: Summative Study V A+B

1 / USER STUDY PLOT – VERSION A



① WELCOME AND INTRODUCTION (1 min)

- +Our prototype uses artificial intelligence to support radiologists in their reporting workflow.
- +The purpose of this study is to measure how you assess the support of artificial intelligence
- +The study is a non-profit project. It is a cooperation between the University of Technology of Tallinn, the Cyprus University of Technology, the LMU Munich and a start-up from Munich called deepc.

+ **Methodology and documentation:**

You will start by using the prototype to solve four tasks. During the study instructions and explanations will be available to you. Afterwards you will answer questionnaires. In the end, we have some questions about your opinion and your demographics.

Due to validation purposes, we need to record your screen during the session (audio and webcam access will not be required) - the screen-recording is needed for backup and documentation only.

② CONSENT FORM (0 min)

- + I understand that my participation is kept confidential in this study, and I am even obliged to maintain confidentiality regarding study content, product information and the used product prototypes. All recordings, personal information and individual results will not be disclosed to third parties.

I understand that I may withdraw at any time from participation in the study. I understand that I may withdraw at any time from participation in the study. I confirm that I agree with the described procedure by selecting the “agree” box below.

③ “HUMAN IN THE LOOP” PROTOTYPE (10 min)

+ **Instructions:**

Before we start: Your tasks will be provided at the beginning of each scenario. You can review the tasks at any time again by clicking the “show” bubble in the bottom left corner.

Feel free to look at the screens carefully and open up the task description after each step so you don't need to memorize it. Please beware: This is a prototype with limited functionality.

1 / USER STUDY PLOT – VERSION A



After you feel you have completed the task, open the “show” bubble and hit the “task completed” button. You will then be redirected to your next task automatically. Let’s try it! You have completed your first task by reading the instructions. Please continue.

- + Task 1 - Start
- + Task 2 - Navigate
- + Task 3 - Findings

④ QUESTIONNAIRE I + II (5 min)

- + Questionnaire: Human Computer Trust Scale (HCTS)
- + Questionnaire: System Usability Scale (SUS)

⑤ INTERVIEW QUESTIONS + DEMOGRAPHICS (5 min)

- + Cool off question: future considerations
- + Demographical questionnaire: personal information

⑤ THANK YOU & GOODBYE

- + Thank you again for providing us your thoughts and to take the time to participate in this study!
If you have any additional feedback, let us know. Have a good day!

3 PROTOTYPE - TASKS



- Please read and use your “Quick Guide” to accomplish the given tasks!
- Not you are testet but the prototype
- The prototype has limited functionality and only specific directions are possible

Task 1 – Start

Imagine yourself sitting in front of your working area. You have just screened a patient. You start the software to review the image set.

1. Log-in to the platform using the provided credentials
2. Select a patient from your list that has not been reviewed yet

Task 2 – Navigate

Now you can view the available data of your patient. The AI has detected some anomalies.

Please select an image where the AI has detected an anomaly.

(Normally you can scroll through the set of images, today you need to click into the specific area).

Task 3 – Findings

Imagine you have studied the image, now you want to document your findings on the anomaly.

1. Draw an arrow in your region of interest
2. Lets assume you see that it is a “A medium localized intraparenchymal hemorrhage in the thalamus-right region.” Please enter your documentation.
3. Finish the process documentation

1 / USER STUDY PLOT – VERSION B



① WELCOME AND INTRODUCTION (1 min)

- +Our prototype uses artificial intelligence to support radiologists in their reporting workflow.
- +The purpose of this study is to measure how you assess the support of artificial intelligence
- +The study is a non-profit project. It is a cooperation between the University of Technology of Tallinn, the Cyprus University of Technology, the LMU Munich and a start-up from Munich called deepc.

+ **Methodology and documentation:**

You will start by using the prototype to solve four tasks. During the study instructions and explanations will be available to you. Afterwards you will answer questionnaires. In the end, we have some questions about your opinion and your demographics.

Due to validation purposes, we need to record your screen during the session (audio and webcam access will not be required) - the screen-recording is needed for backup and documentation only.

② CONSENT FORM (0 min)

- + I understand that my participation is kept confidential in this study, and I am even obliged to maintain confidentiality regarding study content, product information and the used product prototypes. All recordings, personal information and individual results will not be disclosed to third parties.

I understand that I may withdraw at any time from participation in the study. I understand that I may withdraw at any time from participation in the study. I confirm that I agree with the described procedure by selecting the “agree” box below.

③ “HUMAN IN THE LOOP” PROTOTYPE (10 min)

+ **Instructions:**

Before we start: Your tasks will be provided at the beginning of each scenario. You can review the tasks at any time again by clicking the “show” bubble in the bottom left corner.

Feel free to look at the screens carefully and open up the task description after each step so you don't need to memorize it. Please beware: This is a prototype with limited functionality.

1 / USER STUDY PLOT – VERSION B



After you feel you have completed the task, open the “show” bubble and hit the “task completed” button. You will then be redirected to your next task automatically. Let’s try it! You have completed your first task by reading the instructions. Please continue.

- + Task 1 - Start
- + Task 2 - Navigate
- + Task 3 - Findings
- + Task 4 - Impression + Finish

④ QUESTIONNAIRE I + II (5 min)

- + Questionnaire: Human Computer Trust Scale (HCTS)
- + Questionnaire: System Usability Scale (SUS)

⑤ INTERVIEW QUESTIONS + DEMOGRAPHICS (5 min)

- + Cool off questions: personal preferences
- + Demographical questionnaire: personal information

⑤ THANK YOU & GOODBYE

- + Thank you again for providing us your thoughts and to take the time to participate in this study!
If you have any additional feedback, let us know. Have a good day!

3 PROTOTYPE - TASKS



- Please read and use your “Quick Guide” to accomplish the given tasks!
- Not you are testet but the prototype
- The prototype has limited functionality and only specific directions are possible

Task 1 – Start

Imagine yourself sitting in front of your working area. You have just screened a patient. You start the software to review the image set.

1. Log-in to the platform using the provided credentials
2. Select a patient from your list that has not been reviewed yet

Task 2 – Navigate

Now you can view the available data of your patient. The AI has detected some anomalies.

Please select an image where the AI has detected an anomaly.

(Normally you can scroll through the set of images, today you need to click into the specific area).

Task 3 – Findings

Imagine you have studied the image, now you want to document your findings on the anomaly.

1. Draw an arrow in your region of interest
2. Lets assume you see that it is a “medium localized hyperdense finding in the frontal-left region.” Please write your documentation. *(Real typing is not possible but simulated by pressing the “Enter” button).*
3. Finish the process documentation

Task 4 – Impression + Finish

The AI now provides an impression on what it thinks it is seeing in the image and based on the documentation you have just entered.

1. Accept or reject the impression
2. Sign and Approve the report for this finding.

3 QUESTIONNAIRE (HCTS)



YOUR OPINION

1 2 3 4 5

I believe that there could be negative consequences when using deepcOS

Strongly disagree Strongly agree

1 2 3 4 5

I feel I must be cautious when using deepcOS

Strongly disagree Strongly agree

1 2 3 4 5

It is risky to interact with deepcOS

Strongly disagree Strongly agree

1 2 3 4 5

I believe that deepcOS will act in my best interest

Strongly disagree Strongly agree

1 2 3 4 5

I believe that deepcOS will do its best to help me if I need help

Strongly disagree Strongly agree

1 2 3 4 5

I believe that deepcOS is interested in understanding my needs and preferences

Strongly disagree Strongly agree

3 QUESTIONNAIRE (HCTS)



YOUR OPINION

1 2 3 4 5

I think that deepcOS is competent and effective in image interpretation

Strongly disagree Strongly agree

1 2 3 4 5

I think that deepcOS performs its role as supporting software very well

Strongly disagree Strongly agree

1 2 3 4 5

I believe that deepcOS has all the functionalities I would expect from a software that should support image interpretation

Strongly disagree Strongly agree

1 2 3 4 5

If I use deepcOS, I think i would be able to depend on it completely

Strongly disagree Strongly agree

1 2 3 4 5

I can always rely on deepcOS for image interpretation

Strongly disagree Strongly agree

1 2 3 4 5

I can trust the information presented to me by deepcOS

Strongly disagree Strongly agree

3 QUESTIONNAIRE (SUS)



YOUR OPINION

1 2 3 4 5

I think that I would like to use this system frequently.

Strongly disagree Strongly agree

1 2 3 4 5

I found the system unnecessarily complex.

Strongly disagree Strongly agree

1 2 3 4 5

I think that I would need the support of a technical person to be able to use this system.

Strongly disagree Strongly agree

1 2 3 4 5

I found the various functions in this system were well integrated.

Strongly disagree Strongly agree

1 2 3 4 5

I thought there was too much inconsistency in this system.

Strongly disagree Strongly agree

1 2 3 4 5

I would imagine that most people would learn to use this system very quickly.

Strongly disagree Strongly agree

3 QUESTIONNAIRE (SUS)



YOUR OPINION

1 2 3 4 5

I found the system very cumbersome to use.

Strongly disagree Strongly agree

1 2 3 4 5

I felt very confident using the system.

Strongly disagree Strongly agree

1 2 3 4 5

I needed to learn a lot of things before I could get going with this system.

Strongly disagree Strongly agree

4 INTERVIEW - EN



Please record the interview!

Thank you for your patience.

1. What was the most promising feature regarding AI in your eyes?

The AI provides text completion In the End the AI showed what it learned

The AI provides information about its certainty

The AI provides a impression The AI shows me anomalies

2. Was there anything you missed or like us consider in the future?

4 INTERVIEW - EN



 **Please record the interview!**

Thank you for your patience.

1. Was there anything you missed or like us consider in the future?

4 DEMOGRAPHICS



PERSONAL DETAILS

1 2 3 4 5

Own assessment of technical experience (with electronic products):

not experienced at all very experienced

Occupation	Radiologist <input type="checkbox"/>	Medical Student <input type="checkbox"/>	MTRA <input type="checkbox"/>	Other _____
-------------------	---	---	----------------------------------	----------------

Occupation in Years	0-2 <input type="checkbox"/>	2-5 <input type="checkbox"/>	5-10 <input type="checkbox"/>	10 < <input type="checkbox"/>
----------------------------	---------------------------------	---------------------------------	----------------------------------	----------------------------------

Institution	University clinic <input type="checkbox"/>	State-run hospital <input type="checkbox"/>	Private practice <input type="checkbox"/>	Still studying <input type="checkbox"/>	Other _____
--------------------	---	--	--	--	----------------

Gender	Male <input type="checkbox"/>	Female <input type="checkbox"/>	Divers <input type="checkbox"/>
---------------	----------------------------------	------------------------------------	------------------------------------

Age	> 25 <input type="checkbox"/>	25-30 <input type="checkbox"/>	30-40 <input type="checkbox"/>	40-50 <input type="checkbox"/>	50-60 <input type="checkbox"/>	60 < <input type="checkbox"/>
------------	----------------------------------	-----------------------------------	-----------------------------------	-----------------------------------	-----------------------------------	----------------------------------

9.3 Call for Participation (Summative Study)

Remote-Study with the topic: Artificial Intelligence in Radiologists Workflows

„Keeping the Human in the Loop. A Case Study in the Medical Domain, that Should Increase Trust in Workflows Supported by Artificial Intelligence.“

Background of the study

The development of medical image processing has greatly improved in the last few decades. On the one hand, this has resulted in more images being created per individual scan, but also algorithms, based on artificial intelligence have been greatly improved in terms of their applicability and importance in the field of medical imaging. Nevertheless, there is a lack of acceptance of such applications, partly due to insufficient integration into clinical workflows.

This study aims to test a prototype that is intended to support radiologists in their clinical workflow. Utilizing an improved workflow integration, an increase in efficiency through transparency in connection with artificial intelligence is to be achieved. Also, it should be checked whether the solution has a positive effect on the trust and acceptance of such applications.

Study details

Now this new user interface is to be tested with the help of a prototype.

Only a computer / laptop with Chrome or Firefox is required to participate. The study is carried out in English with the help of "Loop11"¹ and takes about 20 minutes. First, a series of tasks is carried out, followed by two short questionnaires.

Just follow this link and you will be directed to the study:

https://www.loop11.com/ui/?l11_uid=76239

Help me to improve the status quo and be part of a new development!

Thank you very much in advance for your help!

Svenja Dittrich

(svenja@idmaster.eu)

¹ Loop11 requires the installation of a browser extension which, of course, can be uninstalled immediately afterwards. This extension guides you through the study and records the study screen as you go. The data is only used to validate the results and will not be published or passed on to third parties.

