Usability Engineering for Medical Devices using Artificial Intelligence and Machine Learning Technology

A Position Paper of DKE UK 811.4

Michael Engler, Prof. Dr. Christian Johner, Martin Krepcke,
Dr.-Ing. Manuel Isaac Martinez Torres, Martin Stangenberg

Abstract

This paper explores how to apply the established IEC 62366-1 usability engineering process for safe and effective use of Artificial Intelligence (AI) and machine learning in medical devices. We identify unique factors specific to AI technology that can impact safe use and integrate these considerations into a comprehensive analysis of IEC 62366-1. This analysis results in a step-by-step guide with practical recommendations for identifying, evaluating, and mitigating use-related risks specific to AI devices.

Our research confirms the effectiveness of the IEC 62366-1 standard for the development of AI-enabled medical devices. However, as a development-focused standard, IEC 62366-1 does not address the post-market phase. This paper advocates for the inclusion of post-market provisions to ensure the usability of AI-enabled devices is maintained throughout their operational lifecycle, even if these provisions are covered by other standards or regulations.

Table of Contents

1 Introduction

1.1 Background

The rapid integration of Artificial Intelligence (AI) and machine learning technology into medical devices offers exciting possibilities for healthcare. However, in an environment where safety of patients is paramount and healthcare professionals rely on medical devices for accurate diagnosis and treatment, these AI-enabled devices present unique challenges for usability engineering.

AI systems differ from traditional software due to their ability to learn and generate insights and recommendations based on data. This introduces an additional layer of complexity, as users need to understand not just how the device functions but also how the AI arrives at its conclusions. This necessitates a deeper understanding of the AI's capabilities and limitations. These factors can significantly impact the safe and effective use of AI-enabled medical devices and require careful consideration from a usability engineering perspective.

Established usability engineering standards like IEC 62366-1 can be applied to AI-enabled medical devices. These standards are designed to be technology-agnostic and prioritize safety through a risk-based approach.

1.2 Purpose

This paper explores how IEC 62366-1 can be used to ensure the safe and effective use of AI-enabled medical devices by addressing the specific human factors challenges they present. It aims to offer practical guidance and examples for addressing these challenges while applying usability engineering principles as outlined in the process standard IEC 62366-1 to the development of medical devices using AI technology.

While the core principles of usability engineering as outlined in IEC 62366-1 are considered technology agnostic, AI introduces specific challenges and considerations that require a tailored approach. This paper highlights these challenges and the potential shifts in user interaction patterns associated with AI technology. It proposes solutions to address these challenges and critically examines the applicability of IEC 62366-1 to AI-enabled medical devices. The paper identifies areas where the standard's provisions might require adaptation or extension to fully encompass the complexities of AI systems.

This paper presents the DKE UK811.4 perspective on human factors considerations related to AI-enabled medical devices. The goal is to facilitate a common understanding amongst standard writers, regulators, and medical device manufacturers of the implications of AI on human factors and the crucial role of use-related safety in the development of AI systems.

1.3 Intended Audience

While this paper is written for organizations developing AI-enabled medical devices, it addresses both usability engineers as well as experts in the AI space. It also targets regulatory authorities, notified bodies, and standardization organizations involved in setting normative and technical standards in both usability engineering and artificial intelligence/machine learning.

1.4 Scope

To establish clear expectations about what this paper covers, the following sections will define its scope. The section will clarify which topics are discussed and which are not. While the out-of-scope section explicitly excludes some important areas associated with AI technology, it doesn't mean they're unimportant in general. However, the authors believe these topics fall outside the scope of use-related safety as defined by the IEC 62366-1 standard.

1.4.1 In Scope

The safety standard IEC 62366-1 defines a usability engineering process to minimize use errors and reduce risks associated with human-device interaction in medical devices. This includes errors that can arise from users entering incorrect information or misinterpreting the device's outputs. Even though AI systems are advanced and are sometimes seen as superior to conventional systems, they can introduce unique use-related risks when users interact with their features. This paper focuses on these specific use-related risks associated with AI functionalities in medical devices.

1.4.2 Out of scope

This section clarifies areas not covered in detail within this paper, as they fall outside the realm of usability engineering for AI medical devices.

- **AI not apparent at the user interface:**

  Not all uses of AI technology will be evident at the device's user interface. Examples include AI algorithms functioning in the background. This means they don't directly interact with the user interface and are not controlled by the user through that user interface.

  As an illustration, implantable devices such as pacemakers or neurostimulators might leverage AI technology. The primary emphasis lies in ensuring the autonomous and secure operation of the device, without necessitating a detailed understanding of the intricacies of the embedded AI. In this case, user awareness of the AI technology is secondary to the device's autonomous and safe operation and a human factors evaluation focusing on AI-related use errors may not be applicable, as the AI functions are transparent to the user, and user involvement is intuitive rather than consciously controlled.

- **AI-related risks outside of usability engineering:**

  The scope of this document intentionally focuses on mitigating AI-related risks within the domains of human factors and usability engineering exclusively. While AI systems may introduce novel risks, not all of them are directly tied to the device use. Risks related to AI-technology not directly related to use can include fairness concerns, data security and privacy risks as well as system reliability. Although the risk management process (ISO 14971, 2019) requires addressing all these risks in medical device design they are not discussed in this paper.

  Furthermore, ethical considerations, legal risks, and societal impacts are often associated with AI technology in medical devices. While unquestionably important aspects, they are beyond the scope of this paper as well.

- **User involvement in AI model development:**

  AI model development often requires the involvement of prospective users in the data labelling process. This labelled data serves as the foundation for training and testing of the AI model. The quality of the data labelling directly impacts the quality of the AI system's performance. Data labelling involves associating labels with the data using user interfaces. While usability engineering plays a crucial role in optimizing data labelling tools, this paper does not address this specific aspect.

- **Human factors considerations unspecific to AI:**

  This document focuses exclusively on AI-specific use-related risks and human factors considerations. Risks and factors unrelated to AI, but inherent in the broader medical device domain, are explicitly excluded. By concentrating on the unique aspects of AI technologies, this paper aims to provide targeted insights and guidance specific to the challenges posed by their integration into medical devices.

- **Malicious use:**

  AI systems introduce novel avenues for potential misuse by malicious actors, a topic extensively covered within the realm of security discussions. IEC 62366-1 categorizes such uses as "Abnormal Use," defining them as activities beyond the reasonable space for human factors engineering. Consequently, the scope of IEC 62366-1 explicitly excludes abnormal use. Similarly, any deliberate misuse or malicious activity involving an AI system is deemed beyond the scope of this document.

2 Terminology

To ensure a clear understanding of the concepts discussed in this paper, the following key terms are defined:

- **AI (artificial intelligence)**
  There are various definitions of artificial intelligence (AI) described in literature. This document does not seek to introduce a new definition or endorse a specific one. Instead, the term 'AI' (artificial intelligence) is used in a general sense to encompass all forms of artificial intelligence, including but not limited to those utilizing machine learning technologies.

- **bias**
  systematic difference in treatment of certain objects, people, or groups in comparison to others
  Note: Treatment is any kind of action, including perception, observation, representation, prediction, or decision. (ISO/IEC 22989, 2022)

- **continuous learning**
  incremental training of an AI system that takes place on an ongoing basis during the operation phase of the AI system life cycle. (ISO/IEC 22989, 2022)
  Note: Continuous learning implies unsupervised learning by the AI system and is one form of machine learning.

- **interpretability**
  the level to which a user gains, and can make use of, both the information embedded within

explanations given by the system and the information provided by the system's transparency level. (Adapted from (Tomsett, Braines, Harborne, Preece, & Chakraborty, 2018): replaced "agent" with "user")

- **explainability**
  the level to which a system can provide clarification for the cause of its decisions/outputs. (Tomsett, Braines, Harborne, Preece, & Chakraborty, 2018)

- **machine learning**
  process of optimizing model parameters through computational techniques, such that the model's behaviour reflects the data or experience. (ISO/IEC 22989, 2022)

- **mental model**
  an individual's internal framework or representation of a concept, system, or situation, shaping how they interpret information and make decisions. (based on (Staggers & Norcio, 1993))

- **transparency**
  the level to which a system provides information about its internal workings or structure, and the data it has been trained with. (Tomsett, Braines, Harborne, Preece, & Chakraborty, 2018)

- **training data**
  data used to train a machine learning model. (ISO/IEC 22989, 2022)

- **use error**
  user action or lack of user action while using the medical device that leads to a different result than that intended by the manufacturer or expected by the user. (IEC 62366-1, 2020)

3 A Model for Analysing Usability in the Context of AI

This section proposes a model to systematically identify and analyse potential usability issues that could lead to use errors with AI-enabled medical devices. This model leverages established frameworks, such as the Perception-Cognition-Action (PCA) model for understanding user behaviour and the AI system functional view introduced by ISO 22989. By integrating these perspectives, the model creates a framework tailored to use error analysis in AI systems.

3.1 PCA Model

In human factors engineering, the Perception, Cognition, and Action (PCA) model serves as a foundational framework for comprehending how individuals interact with technical systems. When considering human factors with AI devices, the PCA model becomes particularly relevant. It guides the exploration of how users perceive information, process it cognitively, and physically interact with these advanced technologies. Refer to Figure 1 for a visual representation of the PCA model, adapted from (Rajan & Redmill, 1996):

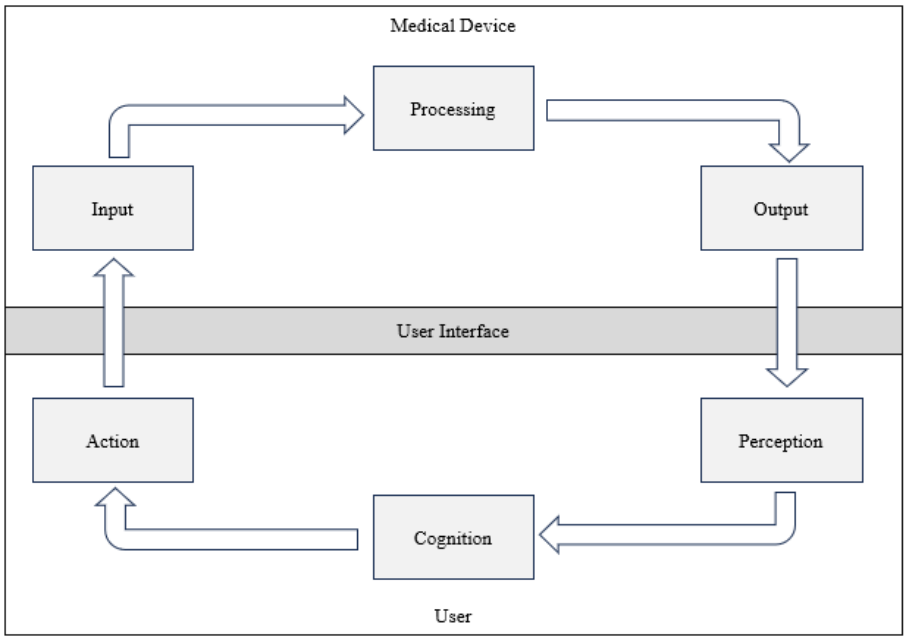


Figure 1 - PCA model.

The PCA model denotes Perception, Cognition, and Action. The PCA model treats the system essentially as a black box hidden behind the user interface, with the user interface being the only perceivable part of the system. The system is categorized into three stages:

- **Input**: Information provided by the user into the system (e.g., entering symptoms into a medical device).
- **Processing**: The internal workings of the system (e.g., the AI algorithm analysing the symptoms).
- **Output**: Any information delivered by the system to the user (e.g., the medical device suggesting potential diagnoses).

Analogously, the user engages in a reciprocal process of Perception, Cognition, and Action.

- **Perception**: Focuses on how individuals gather information from the system through their senses. This includes visual clues, auditory signals, and potentially tactile information. For instance, in a medical device with AI-based image analysis, perception involves the user seeing the system's output on a screen (visual) and potentially hearing audio alerts (auditory).
- **Cognition**: Includes all mental processes like thinking, learning, recollecting, and problem-solving. In the context of usability engineering, it highlights the need for interfaces or systems to align with how users naturally process information. For example, a well-designed AI-enabled medical assistant should present information clearly and logically, reducing the cognitive load on the user and making tasks more intuitive.
- **Action**: Encompasses any physical interaction users have with the system, including pressing buttons, providing input, or any other physical manipulation. In the context of the medical device example, action could involve the user entering information through a keyboard or actuating a physical control, such as a dial.
  It is important to note, that the action does not necessarily have to take place at the AI-device's user interface. For example, a physician might administer a patient's medication through a separate device like a syringe based on recommendations from the AI system.

The PCA model serves as a tool for designers and engineers, facilitating the analysis and optimization of human-system interactions by placing focus on the user's physiological and cognitive abilities.

PCA-Model by Example

Action	A diabetic inserts a glucose test strip into the glucose meter.
Input	Test strip.
Processing	The device measures the glucose level.
Output	The glucose level is displayed on the screen.
Perception	The user reads the glucose level.
Cognition	The user understands the value and estimates the insulin dose.
Action	The user sets the insulin dose on an insulin pen.

Note: Use errors, as defined in IEC 62366-1, occur at the action level, such as setting the wrong insulin dose on the insulin pen.

3.2 AI System Functional View

The ISO/IEC 22989:2023 standard, titled 'Information technology - Artificial intelligence - Artificial intelligence concepts and terminology,' employs a model to illustrate a "*functional view of an AI system, where inputs are processed using a [data] model to produce outputs, and that [data] model can be either built directly or from learning on training data. The parts drawn with dashed lines are for ML [(machine learning)] based AI Systems.*" (ISO/IEC 22989, 2022).

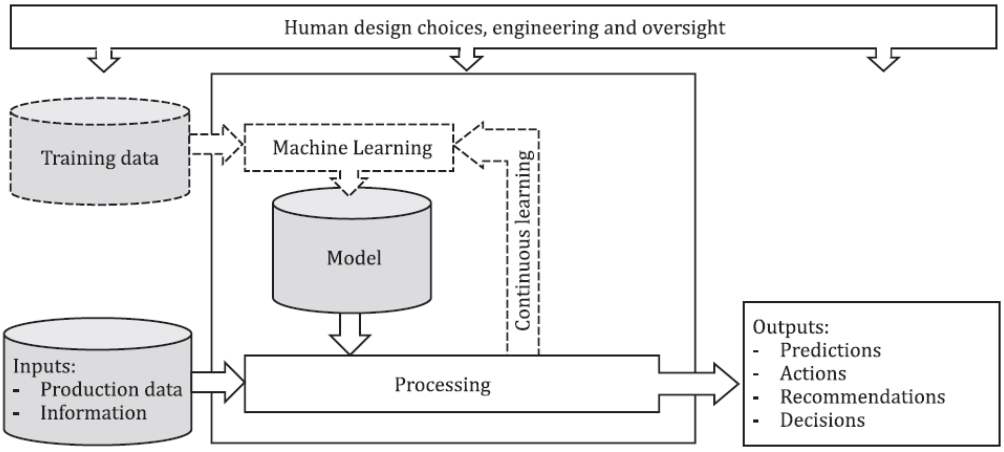


Figure 2 - AI system functional view from ISO/IEC 22989.

This model aligns with the PCA model utilized in usability engineering, using the same categorization of the system into input, processing, and output. In this representation, users are considered one possible source for delivering input and a possible consumer for generated output. However, it's noteworthy that in Figure 2, input and output can also originate from or be directed towards other entities rather than the user. For instance, input may come from environmental data obtained through sensors, and output may be directed to other systems or actuators. In such cases where sensors provide input, the user is likely not directly involved in providing that data.

3.3 Extended PCA Model for AI Systems

Usability engineering traditionally focuses on scenarios where users provide input to a system and receive output from it. By incorporating the user into the AI system functional view, an extended model that merges the AI system's functional view with the PCA model can be created. This extended model, illustrated in Figure 3, is the conceptual framework for the discussions within this paper.

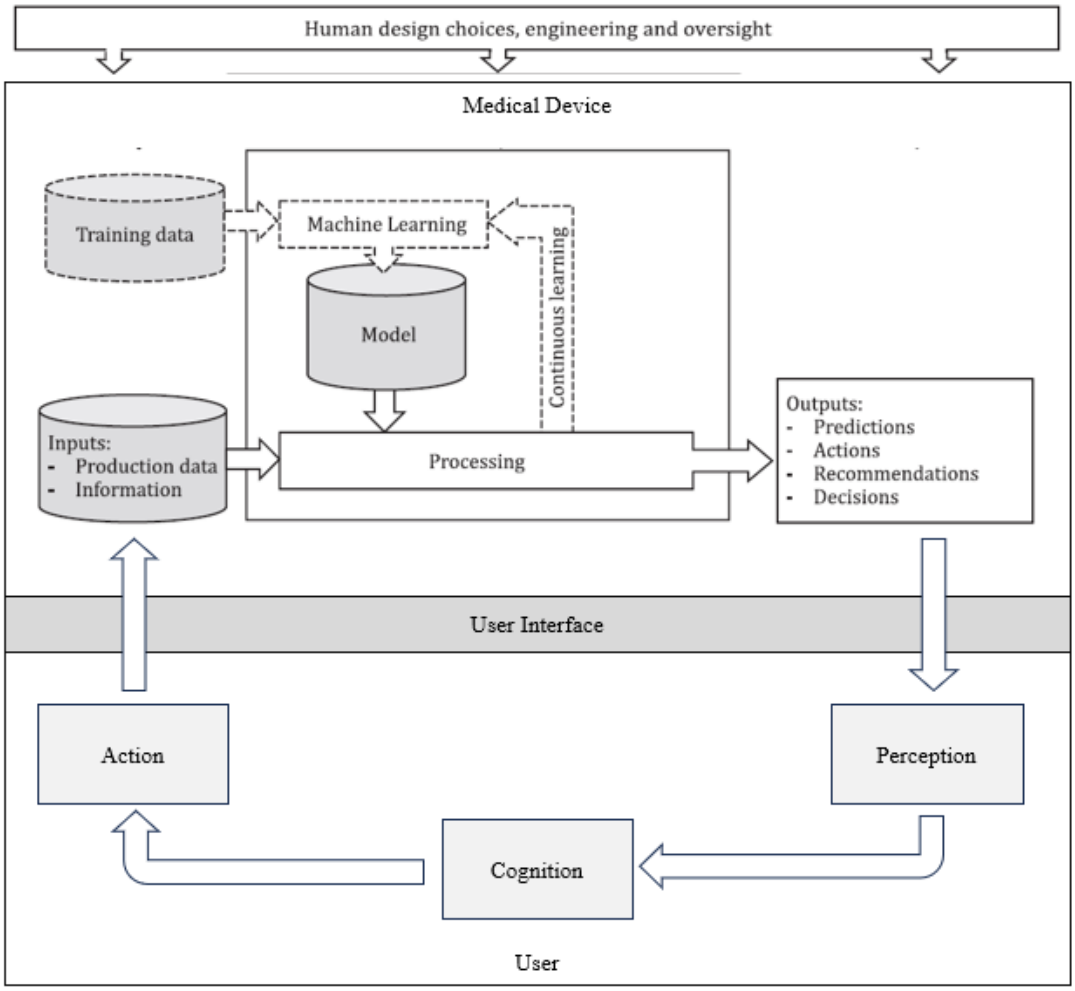


Figure 3 - Extended PCA model.

The extended model indicates two distinct areas of human involvement in AI systems:

- **System Use (Lower Part)**: During system use, users interact with the user interface to achieve specific goals in a real-world context. This is the traditional focus of usability engineering, and it will be the primary concern throughout this paper.
- **System Design and Development (Upper Part)**: Human involvement occurs during the design process. This includes decisions, engineering efforts, and oversight[1]. While crucial, these activities fall outside the scope of traditional usability engineering, even though they might require the involvement of representative users (e.g., radiologists for labelling medical image data) during training and oversight. These activities, while involving real-world users, are considered part of broader system design and design validation rather than user interface evaluations in the stricter sense.

The usability engineering process focuses on the lower portion of Figure 3, analysing the user's behaviour while interacting with the system's user interface to accomplish their goals within their context of use. This

[1] ISO/IEC 22989 points out that the "degree of oversight depends on the use case. At a minimum, oversight is typically present during training and validation."

focus will be maintained throughout the remainder of this document. Additionally, this document addresses risks through AI-specific interaction patterns that need different approaches to usability engineering.

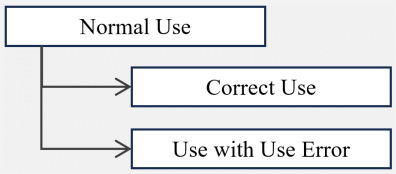
4 Contributing Factors to Use Error in AI-enabled Medical Devices

The previous chapter has established a theoretical foundation for understanding how users interact with medical devices through the PCA (Perception, Cognition, Action) model. This model is particularly useful for identifying sequences of events that may lead to use errors, delineating how perceptual issues can influence cognitive processes and ultimately result in actions that deviate from correct use. Within the domain of AI, this framework is further refined to account for the unique characteristics of AI systems and the specific contexts in which they operate.

This section applies the PCA model's logic to identify factors contributing to use errors commonly associated with AI technology. Figure 4 illustrates the relationship between system properties, context of use, and user interaction stages. For example, a complex user interface may cause information overload (Perception), leading to misinterpretation of AI outputs (Cognition), and finally resulting in an erroneous action (Use error). It is important to recognize that the factors listed here do not represent an exhaustive list but should be seen as a starting point for the analysis of a specific AI-enabled medical device. In any case, the structured approach presented in this section can be used for systematically analysing potential use errors within the realm of AI-enabled medical devices.

Understanding Use in Medical Devices

The diagram below illustrates the relationship between Normal Use, Correct Use, and Use Error:



Normal Use: All actions a user might take when operating a device as intended, including potential use errors.

Correct Use: Operating the device following the instructions for use without errors. This is a subset of Normal Use.

Use with Use Error: Operating the device but committing a Use Error. For example, a user's action leads to an outcome not intended by the user. Use errors are considered part of Normal Use.

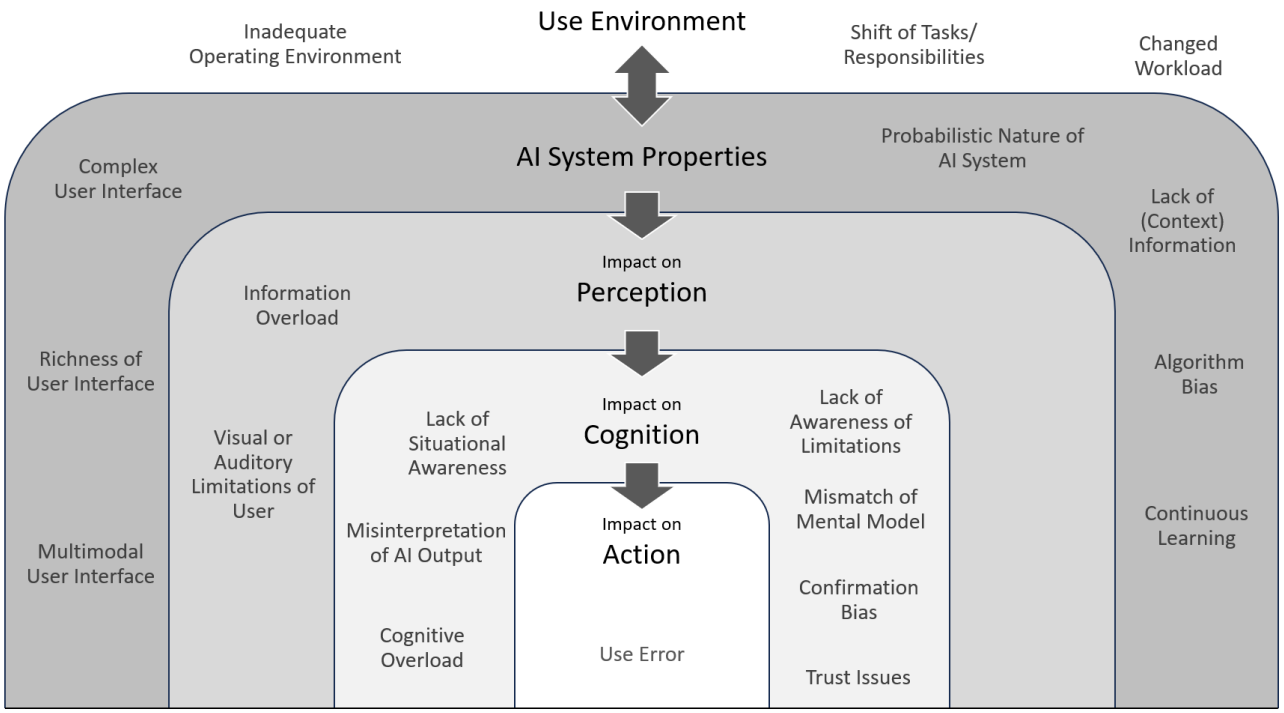


Figure 4 - Contributing factors to use error in the context of AI technology.

4.1 Use Environment

The integration of AI technologies into medical devices can significantly alter the context in which these devices operate. Although designed to support their users, AI systems can adversely impact the workload of healthcare professionals due to novel uses of the device. Understanding these changes is crucial for identifying potential use errors and designing user interfaces that mitigate these risks. The table below details the specific aspects of the use environment that are particularly influenced by AI technologies.

Factor	Description	Example
Shift of Tasks/Responsibilities	AI systems can redistribute or redefine tasks and responsibilities and can alter traditional patterns of cooperation among healthcare professionals. These changes often require the development of new skills and may even lead to devaluating existing skillsets.	An AI diagnostic tool might take over preliminary data analysis, shifting the physician's focus from routine tasks to interpreting complex cases that the AI flags as ambiguous.
Changed Workload	The adoption of AI technologies can both increase and decrease the cognitive and physical workload on users, depending on how tasks are automated or augmented.	While AI can reduce the manual workload by automating data entry, it may increase cognitive workload by requiring users to interpret more complex data or manage additional oversight tasks.
Inadequate Operating Environment	The effectiveness of AI systems can be compromised by operating environments that were not adequately considered during design, such as those with poor lighting, high noise levels, or other distractions.	A voice-activated medical recording system struggles in environments with high background noise, leading to frequent misrecognitions and errors in patient records.

Table 1 - Contributing Factors of Use Environment

4.2 AI System Properties

4.2.1 Properties of AI Technology

The integration of AI technology in medical devices not only enhances capabilities but also introduces complex challenges. This section explores the probabilistic nature of AI systems, the impact of lacking context information, concerns about algorithm bias, and the implications of continuous learning, as these factors can significantly influence device usability and effectiveness. Moreover, the probabilistic operations of AI, coupled with a degree of non-transparency from the user's perspective, often described as a 'black box' phenomenon, create unique challenges. This combination can obscure the rationale behind system outputs, making it difficult for users to understand and trust the decisions made by AI-enabled devices.

Factor	Description	Example
Probabilistic Nature of AI System	AI systems often produce results based on probabilities, where even slight variations in input can change output substantially due to the uncertainty inherent in the model's predictions.	A diagnostic AI analysing the exact same set of patient data on two different occasions might yield varying risk assessments for a condition, despite no change in the patient's information. This variability can introduce challenges in clinical decision-making, as it reflects

Factor	Description	Example
		the system's probabilistic approach rather than a definitive diagnostic outcome.
Lack of (Context) Information	AI systems may not always have access to or incorporate necessary context information, impacting the assessment of output correctness.	For instance, a telemedicine AI system used by a physician might lack crucial clinical information obtainable through direct patient contact. This absence of context could lead to misinterpretation of symptoms or oversight of critical health indicators, affecting the quality of care provided.
Algorithm Bias	Despite efforts to minimize bias, AI algorithms can exhibit preferences or inaccuracies due to the data they were trained on, potentially affecting outcomes when applied to diverse populations.	An AI system trained predominantly on data from one demographic might be less accurate for patients outside that group, inadvertently leading to less optimal care recommendations. Users might inadvertently use the system on patients for whom the system was not specifically trained.
Continuous Learning	AI systems that learn and adapt continuously can produce varying results for similar inputs over time, complicating consistency in diagnostics or treatment recommendations.	An AI-enabled medication management system initially provides recommendations for drug dosages based on a patient's medical history and current health data. As the system continues to learn from a broader pool of patient data over time, it might adjust its dosage recommendations. This continuous learning process could lead to variations in suggested dosages for patients with similar health profiles, complicating the task of maintaining consistent treatment protocols.

Table 2 - Contributing Factors of AI System Characteristics

4.2.2 Properties of AI-enabled User Interfaces

The integration of AI technologies in medical devices introduces a level of complexity in user interfaces and interactions that is unprecedented. This section explores the impact of complex, rich, and multi-modal interfaces on user experience, particularly emphasizing challenges introduced by AI functionalities.

Factor	Description	Example
Complex user interface	AI systems introduce new types of complex inputs and outputs, like natural language processing and 3D visualizations, complicating navigation and functionality.	An AI diagnostic tool utilizing natural language inputs for symptom description and providing diagnosis through complex 3D models of affected areas can overwhelm users not versed in interpreting detailed visual data, potentially leading to overconfidence in results.

Factor	Description	Example
Richness of user interface	The capability of AI to process and display extensive information, from statistics to metaverse integrations, adds a layer of complexity by presenting a wealth of data simultaneously.	A clinical decision support system using AI to analyse patient data presents statistics, treatment options, and prognostic predictions all at once, requiring clinicians to sift through a dense array of information to make informed decisions.
Multi-modal user interface	AI technologies enable multimodal interactions allowing flexible input (speech, handwriting, gestures) and output (speech synthesis, graphics), which can simplify, but also complicate user experience.	A rehabilitation device using voice commands, haptic feedback, and visual progress graphs can be problematic for elderly patients who may struggle with voice recognition or interpreting complex visual data.

Table 3 - Contributing Factors of AI-enabled User Interfaces

4.3 Impact on Perception

The perception level focuses on the initial sensory interaction users have with AI-enabled medical devices, emphasizing the role of sight, sound, and touch. The introduction of novel user interfaces in AI-enabled medical devices aim to improve usability by offering more intuitive interaction methods and providing comprehensive data in accessible formats. However, if these interfaces are poorly designed, they can introduce adverse elements that hinder effective perception, for example through excessive information presented to the user. The perception stage is critical because it sets the foundation for subsequent cognitive processing and action. Poorly designed interfaces can inadvertently lead to sensory overload or miss the unique sensory limitations of diverse users. The following table displays challenges at the perception level posed by AI interfaces.

Factor	Description	Example
Information Overload	An excess of information presented through the interface can overwhelm the user's sensory capacity, leading to important details being missed.	A clinician facing multiple alerts and recommendations from an AI system while trying to prioritize patient care tasks, potentially overlooking critical alerts.
Visual or Auditory Limitations of User	Interfaces that do not account for the user’s visual or auditory limitations may fail to effectively communicate information, leading to use errors.	A voice-activated virtual health assistant may not be effectively used by individuals with hearing impairments, or detailed health data charts may be inaccessible to users with visual impairments, resulting in missed or misunderstood health information.

Table 4 - Contributing Factors of Perception

4.4 Impact on cognition

At the cognition level, the way users process and understand information from AI-enabled medical devices becomes crucial. The complexity of these devices, coupled with the novel user interfaces and the vast amount of data they provide, can significantly impact cognitive processes. The following table provides a list of various cognitive challenges that arise when interacting with AI technologies.

Factor	Description	Example
Cognitive Overload	The volume and complexity of information can overwhelm users, hindering their ability to process and make decisions.	A medical AI system designed for patient treatment planning requires users to input a wide range of patient data, interpret AI-generated treatment options, and consider probabilistic outcomes for each option. The need to synthesize this information, alongside understanding the AI's reasoning process for each option, places a significant cognitive burden on healthcare providers, potentially leading to decision fatigue or errors in treatment selection.
Mismatch of Mental Model	Users may struggle to integrate AI-generated results with their clinical knowledge, leading to confusion.	A doctor finds it difficult to reconcile AI-generated diagnosis suggestions with their clinical experience, potentially doubting accurate AI recommendations.
Misinterpretation of AI Output	Users may misunderstand the complex outputs generated by AI, especially when lacking context or explanations.	A healthcare provider misinterprets a probabilistic prediction of disease risk from an AI tool as a definite diagnosis, leading to unnecessary anxiety or incorrect patient counselling.
Lack of Awareness of Limitations	Users might not be fully aware of the AI system's limitations, including the accuracy and appropriateness of its outputs.	A physician overestimates the capability of an AI diagnostic tool, not realizing it has limitations in recognizing rare conditions not covered in its training data.
Lack of Situational Awareness	Users may not fully grasp the broader context or environment in which the AI system operates, potentially leading to inappropriate reliance on AI decisions.	A doctor focuses solely on an AI-suggested diagnosis without considering the patient's recent symptoms or ongoing medications, potentially overlooking a crucial contributing factor.
Confirmation Bias	Users may favour information that confirms their pre-existing beliefs, disregarding AI outputs that contradict them.	A specialist ignores an AI recommendation for a less common treatment path, favouring a more familiar approach despite AI data suggesting better outcomes with the alternative. A specialist suspects a particular condition of a patient. However, the AI system suggests a different diagnosis. Instead of reviewing all the patient's information carefully, the specialist

Factor	Description	Example
		focuses only on details that seem to confirm their initial suspicion. This selective review reinforces the specialist's (potentially wrong) belief.
Trust Issues - Overtrust	Users may place too much trust in the AI system, leading to overreliance. This can lead to false decision making. Automation bias is one form of overtrust.	Overtrust: A nurse disregards manual double-checking of medication dosages, relying solely on the AI system's calculations (automation bias).
Trust Issues - Mistrust	Users may place too little trust in the AI system, resulting in underuse. This can lead to false decision making.	Mistrust: A nurse, disregarding the AI's recommended treatment plan for a time-sensitive case due to personal doubts about its accuracy, delays potentially life-saving interventions.

Table 5 - Contributing Factors of Cognition

4.5 Impact on Action

At the action level, the focus shifts to the physical interactions users have with medical devices, which include inputs to the AI system or to another system, operated by the user based on information processed from the AI system. Missteps in these interactions are identified as use errors which can manifest in various forms. While there are multiple ways to categorize use errors, a helpful approach is to consider three primary types: errors in executing actions, delays in action, and the absence of action when necessary. Each type stems from a range of factors, including but not limited to cognitive challenges encountered during device interaction.

Factor	Description	Example
Errors in Executing Actions	Physical actions taken by the user that are incorrect due to perception issues or cognitive issues such as misunderstanding AI outputs or system limitations.	A clinician erroneously subjects a patient to an unnecessary biopsy due to a misunderstanding of the AI-based diagnostic tool's outputs.
Delay in Action	Hesitation or postponement of necessary actions, often resulting from uncertainty or mistrust in the AI system's recommendations.	A doctor hesitates to follow an AI system's treatment recommendation for a patient due to uncertainty about the AI's decision-making process, delaying critical care.
No Action	Failure to take action when it is warranted, possibly due to overtrust in the AI system's autonomous capabilities or a lack of	A doctor reviewing a patient's MRI scans receives AI-generated analysis highlighting a low-risk finding. Trusting the AI's assessment, the doctor fails to order a biopsy, potentially missing a hidden malignancy.

Factor	Description	Example
	awareness of the system's limitations.	

Table 6 - Contributing Factors of Action

5 Usability Engineering Process Best Practices

5.1 Prepare Use Specification

The usability engineering process, as outlined in IEC 62366-1, begins with creating a use specification. The use specification summarizes findings from user research about the intended medical indication, patient and user populations, and the use environment, collectively often called the 'context of use.' This information is crucial for designing medical devices and shapes the design of AI algorithms and their data models.

Particular attention is warranted if the introduction of AI technology fundamentally changes how users interact with the device. When AI technology significantly deviates from conventional systems and has the potential to reshape medical practice, it is important to address these paradigm shifts. Some AI applications can completely redefine jobs, tasks, and responsibilities in the medical field, necessitating careful consideration during the usability engineering process.

The following sections explore specific considerations for creating a use specification for medical devices that use AI technology.

5.1.1 Intended Medical Indication

A clear definition of the medical conditions a device aims to address is crucial for successful AI development. It provides a foundation for the development team, outlining the specific patient conditions and medical challenges the AI technology will help tackle. Understanding the medical context in depth is essential for designing and implementing effective AI algorithms.

For instance, consider a medical device designed to assist in the diagnosis of neurological disorders. Beyond the technical aspects, a deep understanding of neurology is crucial to navigate the complexities of the data involved in the diagnostic process. This includes not only recognizing the types of data used but also understanding the subtle variations and patterns indicative of neurological conditions to correctly discriminate neurological disorder from normal neurological function. This depth of medical knowledge is essential for developing accurate and clinically meaningful AI algorithms. Without it, creating algorithms that work effectively in real-world applications becomes challenging and may compromise the device's overall efficacy.

It's important to distinguish between the medical indication and the technical specifications of the AI system. The medical indication focuses on the "what" - the patient conditions the device aims to address. Technical specifications, on the other hand, focus on the "how" - the technical details of how the AI system will be built and implemented.

5.1.2 Intended Patient Population

Understanding the patient population is crucial for effective AI system development. For example, for diagnostic algorithms, it's essential to gather training data that comprehensively represent the diversity within the intended patient group. Failing to do so may introduce biases, impacting the algorithm's effectiveness.

For example, if the device is designed to detect skin conditions using AI, the training data must include a representative sample of various skin tones, ages, and conditions of the intended patient population to avoid biased outcomes. Collaborating closely with healthcare providers is recommended to ensure the AI algorithm is trained on diverse datasets that reflect the characteristics of the patient population it shall serve. This not only enhances the algorithm's diagnostic accuracy but also fosters equity and fairness in healthcare outcomes across different demographic groups.

5.1.3 Intended Part of the Body or Type of Tissue

Understanding the specific part of the body or type of tissue that an AI-enabled medical device interacts with is crucial in its development. Anatomical knowledge of these body segments is fundamental, influencing both the usability and effectiveness of the device.

In the realm of AI-enabled medical devices, considerations specific to the interaction with certain body parts or tissues play an important role in enhancing functionality and precision. For example, automated segmentation of medical images requires tailored algorithms, as different tissues might need distinct segmentation strategies, depending on the imaging modality.

In another scenario, an AI-enabled powered screwdriver is tasked to automatically insert bone screws to the optimal torque. Here, precise anatomical knowledge of bony structures is vital for ensuring the accuracy and safety of the insertion process.

5.1.4 Intended User Profile

Comprehensive knowledge about the intended user population is summarized in a set of user profiles. Different user profiles may be necessary, reflecting distinct roles users may have when operating the medical device. IEC 62366-1 categorizes these roles as user groups, such as healthcare professionals, caregivers, and patients.

Researching the intended user population is crucial in determining if a particular AI technology aligns with the prospective users. For example, an AI support system requiring critical review of its results might not be suitable for emergency care first responders who must make rapid decisions. In such cases, the AI support system needs to produce highly accurate and easily understandable results without room for interpretation.

The integration of AI also may transform the way healthcare providers operate, potentially leading to the establishment of entirely new job profiles and distinct user populations. Instead of a radiologist directly diagnosing images, a specialist aided by an AI system may now review AI-generated results, requiring a different skill set and level of education.

According to Tomsett et al. (Tomsett, Braines, Harborne, Preece, & Chakraborty, 2018) users of an AI system can be categorized as Operators and Executors. Operators interact directly with the system, providing input and receiving its outputs. Conversely, Executors make decisions informed by the AI's output, often receiving information from Operators. For instance, in an AI-based radiology system, the radiologist (Operator)

analyses patient scans and receives the AI's findings (e.g., potential tumours). The surgeon (Executor) leverages this information, alongside their expertise, to make informed surgical decisions. Considering both roles is crucial when defining user profiles, as use errors can happen during direct interaction or when interpreting AI output.

Specified user profiles serve as a tool for the product development team, providing insights to the unique needs, preferences, and proficiency levels, physical abilities, cognitive abilities, intellectual abilities, and expressive abilities of different user groups, especially when dealing with advanced technology. For example, a medical device for elderly lay users demands a significantly different user interface design than devices tailored for highly specialized radiologists, especially when incorporating novel technology such as AI. The introduction of AI may require a reassessment of user interface elements, considering factors like ease of use, interpretability of AI-generated results, and accommodation of diverse skill levels.

User profiles are also useful to assess the training and skill levels required for effective device operation. Using the example of the AI-based diagnostic system mentioned earlier, specific training for healthcare professionals may be necessary to interpret AI-generated results, particularly in discerning when to trust or mistrust them, in particular for systems that learn over time during operation.

5.1.5 Intended Use Environment

The intended use environment for an AI-enabled medical device can vary, including places like hospitals, homes, or clinics. Environmental factors can impact not just the device's usability but also the effectiveness of its AI algorithms or models.

Identifying the physical and social environments in which users expect to interact with the medical device is essential. For instance, a device utilizing a novel AI-based input modality like speech recognition might not be suitable for public spaces, where users might feel uncomfortable if others can overhear their interactions with the system. Similarly, a system providing auditive outputs in natural language may pose challenges in certain settings due to data privacy concerns.

The physical characteristics of the use environment, such as noise levels and lighting conditions, can also impact the effectiveness of the AI algorithm which might depend on those environmental conditions. Designing the device to adapt to different environmental conditions may be necessary for optimal functionality. For example, a device controlled through speech recognition intended for use in both noisy and quiet environments, requires an adaptive algorithm capable of performing well in both situations.

The introduction of AI can radically change use environments. For example, a clinician or physician might no longer work in a hospital but remotely, or the patient might be remote and not physically present. In some scenarios, clinicians may work in settings like radiology centres, where they continuously review and supervise the outputs of AI systems in a 'call centre-like' environment. This represents a significant shift from current practices, transforming the social context of their work. For instance, instead of collaborating directly with other healthcare professionals and interacting within a hospital setting, they may now work in isolation or in a highly structured, repetitive task environment, impacting teamwork dynamics and job satisfaction. Therefore, in the context of AI, it is not enough to understand and consider the current use environment. For successful user interface design, it is crucial to anticipate and consider the prospective use environment, as this is the context in which users will operate the newly developed device.

5.1.6 Operating Principle

The operating principle is the only element of the use specification that does not address the context of use (medical indication, patients, users, and the use environment), but focuses on the primary operational mechanisms of the envisioned medical device. Hence, the operating principle shifts from the problem domain (context of use) to the solution domain (medical device/system).

The operating principle offers a chance for a critical discussion, assessing how the primary operations of the medical device align with the previously described context of use. This becomes important, especially when dealing with innovative technologies like AI-enabled medical devices. For an AI-enabled medical device, the operating principle should cover various aspects related to user interaction, including, where applicable:

- Model architecture or type of AI algorithm (e.g., Large Language Model, Classification, Regression),
- Characteristics of training data,
- Input and output modalities of the user interface,
- Means to establish explainability of the AI system,
- Capability of continuous learning, especially if user involvement is planned,
- Mechanisms of monitoring of the AI system's performance, particularly when it involves user interaction.

While a comprehensive operating principle may not be fully attainable in the early stages of development, it should evolve and be refined as the project progresses and more information becomes available. Describing a thorough operating principle early in the project allows the development team to use it as a foundation for subsequent analyses, including the identification of use-related risks associated with introducing this new medical device technology into the given context of use and strategies to mitigate these risks.

5.2 Identify User Interface Characteristics related to Safety and potential Use Errors

IEC 62366-1, Clause 5.2, mandates that manufacturers identify user interface (UI) characteristics that can impact safe use and contribute to potential use errors. This analysis involves evaluating factors that might lead to such errors in AI-enabled medical devices.

The specific system properties of AI-enabled devices, as identified in Section 4, can be interpreted as UI characteristics related to safety in the context of this standard. This includes the very nature of AI technology itself. However, it's crucial to consider all other contributing factors outlined in Section 4 as well. These factors can stem from the context of use and the user's perception and cognitive capabilities.

While Section 4 offers a broad range of factors commonly associated with AI technology, it may not be exhaustive. Given the diverse applications of AI in medical devices, additional factors specific to the medical device could also lead to use error. Therefore, manufacturers are encouraged to use the list in Section 4 as a starting point for conducting a more detailed and device-specific analysis of potential use errors.

Derived from section 4 the following table shows some examples of contributing factors relevant to AI technology and associated potential use errors:

Contributing Factors & Examples	Potential Use Error
Shift of tasks and responsibilities: The new workflow imposed by the AI system confuses users.	User misses crucial steps during diagnosis due to a new, automated workflow in the AI system. (Action not performed at all)
Information overload: An AI-based alarm system continuously issues false positive alarms.	User ignores or improperly responds to warnings or alerts generated by the AI system due to desensitization or misunderstanding. (Action not performed at all)
Information Overload: The AI system displays excessive data on the screen.	Physician overlooks a critical alert from the AI system and fails to act appropriately. (Action not performed at all)
Complexity of User Interface: Users not familiar with the complex user interface might feel overwhelmed and experience information overload and cognitive overload.	User enters inaccurate or incomplete data into the AI system, due to the complexity or poor design of the input interface. (Action performed incorrectly)
Complexity of User Interface: An AI system provides high quality patient diagnoses but has confusing menus, for entering the related patient information.	Clinician enters incorrect patient data due to a confusing layout in the AI system's input interface. (Action performed incorrectly)
Lack of Context Information: A personalized fitness app recommends a high-intensity workout routine based on a user's recent activity data. However, the AI doesn't have access to the user's medical history, which includes a recent injury.	The user, trusting the app's recommendation and unaware of the missing context, attempts the strenuous workout. This could lead to reinjury or further health complications. (Action performed incorrectly)
Mismatch in Mental Model: A new insulin pump utilizes an AI algorithm to automatically adjust insulin delivery based on continuous glucose monitoring (CGM) data. However, the algorithm prioritizes maintaining blood sugar within a tight range to minimize short-term fluctuations.	The patient, used to manually adjusting insulin based on a broader understanding of their body's response, experiences frequent pump adjustments and doesn't understand why. This lack of transparency can lead to frustration, decreased trust in the device, and potentially ignoring alerts or malfunctioning of the pump. (Action not performed at all)
Lack of Awareness of Limitations: Users are unaware of the specific limitations of the AI system, leading to misuse or overreliance on its outputs.	Clinician relies on an AI system for skin cancer diagnosis, unaware that the system's predictions are less accurate for patients with darker skin tones, and incorrectly diagnoses a benign mole as malignant, leading to unnecessary biopsy. (Action performed incorrectly)

Contributing Factors & Examples	Potential Use Error
Misinterpretation of AI Output: Seeing a 70% chance of a specific disease on the AI report, a doctor mistakenly interprets it as a confirmed diagnosis.	Doctor misinterprets a probabilistic prediction of disease risk from the AI as a definite diagnosis, leading to unnecessary treatment. (Action performed incorrectly)
Confirmation Bias: Despite the AI suggesting a potentially more effective but less frequently used treatment for a rare cancer, a specialist clings to their preferred, familiar approach.	Specialist disregards an AI recommendation for a less common treatment path, favouring a more familiar approach despite AI data suggesting better outcomes. (Action not performed at all)
Trust Issues – Overtrust: User overtrusts the AI system's recommendations without critical evaluation	Incorrect patient treatment decisions made based on AI recommendations. (Action performed incorrectly)
Trust Issues – Overtrust: Nurse places full trust in the AI system's dosage calculations, neglecting the fact that the system might be wrong.	Nurse fails to perform a double-check on medication dosages, solely relying on the AI system's calculations. (Action not performed at all)

Table 7 - Examples of contributing factors relevant to AI technology and associated potential use errors.

To systematically elicit characteristics related to safety and potential use errors, manufacturers can employ several well-established usability engineering methods, ensuring a thorough examination of how users interact with AI-enabled devices and identifying areas prone to error. These methods include:

- **Context Scenarios**: Traditional usability engineering methods, such as contextual inquiry, are important tools to understand the user's context of use but might fall short for eliciting user tasks and workflows in the context of new innovative technology. Contextual inquiry typically generates current-state scenarios rather than desired future-state scenarios. Prospective users can usually accurately describe their current work environment, tasks, and responsibilities, but they often struggle to predict how these aspects might change with the introduction of new technologies, such as AI. Therefore, it is important to employ methods that focus on future-state scenarios, such as Context Scenarios as described in Cooper et al. (Cooper, Reimann, Cronin, & Noessel, 2014), to better anticipate and design for the changes introduced by AI.
- **Task Analysis**: This involves a detailed examination of each task that users are expected to perform with the device, focusing particularly on areas where AI behaviour could lead to confusion or errors. In the area of AI many usability challenges seem to relate to cognitive tasks. Therefore, a significant emphasis should be placed on understanding the cognitive workload and potential for use errors caused by cognitive errors.
- **PCA-Analysis**: Perception-Cognition-Action (PCA) analysis helps in understanding how users interact with AI systems by breaking down the interaction into three stages: perception, cognition, and action. As seen in section 4 many use errors in AI systems are caused by problems in cognition, such as misunderstanding AI outputs or difficulty processing complex information. By analysing the

P, C, and A stages, manufacturers can identify potential points of failure and design user interfaces that mitigate cognitive overload, misinterpretation, and incorrect actions (i.e., use errors).

- **Prototype Testing/Usability Testing**: Early and iterative testing with device prototypes is crucial for identifying use errors in realistic scenarios. This approach is particularly valuable for examining how users interact with AI features, including the handling of edge cases. By observing user interactions in a controlled environment, developers can pinpoint and mitigate potential errors before they affect safety.
- **Interviews and Focus Groups**: Engaging directly with end-users through interviews and focus groups offers invaluable insights into user expectations, potential misunderstandings, and scenarios of misuse. This direct feedback can illuminate areas of the user interface that may not be as intuitive as intended or where additional guidance is needed to prevent misuse.

5.3 Identify known or foreseeable Hazards and Hazardous Situations

This section focuses on identifying hazards and hazardous situations arising from use errors identified in the preceding section. The objective of the IEC 62366-1 Clause 5.3 is to systematically recognize all potential hazards and hazardous situations associated with AI device user interfaces.

For this discussion the following key definitions are important to understand:

- **Use Error**: An incorrect user input or action stemming from misunderstandings, misinterpretations, or slips during interaction with the device.
- **Hazard**: The final event in a sequence leading to potential harm.
- **Hazardous Situation**: When a person is exposed to a hazard, creating the potential for harm.

These concepts apply equally to both AI and conventional medical devices. However, the inherent complexity and novelty of AI technology introduce additional challenges.

Following the cyclical nature of the PCA model (described in Section 3), IEC 62366-2 outlines two potential pathways through which a use error can lead to a hazardous situation. Figure 5 (adapted from IEC 62366-1) illustrates these cause-effect relationships.

- **Path A: Hazardous Situation Caused by a Device Response**:
  Here, the user's incorrect input leads the device to produce an output that directly creates a hazardous situation without further user involvement. Analysing use errors within this path involves tracing the connection between the use error, the device's internal workings (input, processing, output), and the resulting hazardous situation (A).
- **Path B: Hazardous Situation Caused by Use Error (on Patient or Different Device):**
  In this path, the user misinterprets or fails to act appropriately on the AI device's output. This can lead to a use error within the broader use environment (on the patient or with a different medical device) and ultimately create a hazardous situation (B).

For a comprehensive assessment of potential hazards and hazardous situations, it's crucial to explore both paths. Analysing both scenarios helps to identify a broader range of risks associated with AI-enabled medical devices.

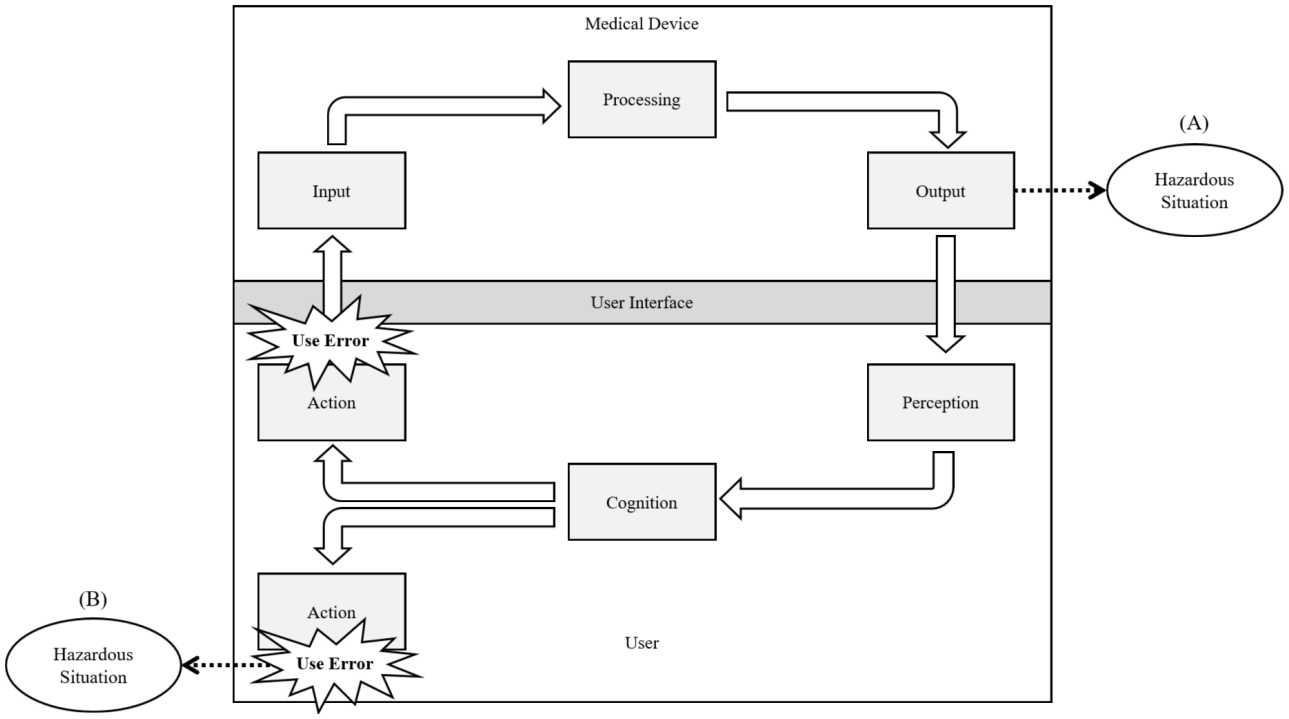


Figure 5 - Cause-Effect relationship between Use Error and Hazardous Situations

The following two tables show examples analysing both paths:

- **Table 8 - Path A - How a use error at** the AI system's user interface can trigger the AI to generate an output that directly leads to a hazardous situation **– Path A**: This table showcases how a use error at the AI system's user interface can trigger the AI to generate an output that directly leads to a hazardous situation.
- **Table 9 – Path B**: This table demonstrates how a user's misinterpretation of the AI system's output can lead to a use error and ultimately a hazardous situation. Although triggered by output of the AI system the hazards and hazardous situations appear in the use environment outside of the AI system

Use Error (at AI system's UI input)	**Hazard**	**Hazardous Situation**
Clinician enters incorrect patient data due to a confusing layout in the AI system's input interface.	AI system calculates incorrect dosage values (based on incorrect patient data).	Patient exposed to an overdose or underdose.
A physician fails to enter crucial clinical information obtained through direct patient contact into the AI system, leaving the system with incomplete information.	AI system misinterprets symptoms or oversees critical health indicators and in consequence creates lower-level quality recommendations.	Patient receives an inadequate treatment plan.

Table 8 - Path A - How a use error at the AI system's user interface can trigger the AI to generate an output that directly leads to a hazardous situation.

Sequence of Events (leading to use error)	**Use Error** (tiggered by AI system's UI output)	**Hazard**	**Hazardous Situation**
Due to the probabilistic nature the AI system outputs inaccurate information. The user, unaware of the system limitation, trusts the inaccurate information.	User administers the wrong medication dosage to the infusion pump.	Wrong medication dosage in infusion pump.	The patient receives an overdose or underdose.
Due to lack of interpretability outputs of the AI-system lack clarity the user misunderstands the AI's decision-making process.	User ignores critical outputs.	Critical output remains unrecognized.	Patient does not receive the necessary treatment.
Due to continuous learning the system behaviour appears inconsistent the user is confused and misinterprets the device's output.	User does not follow the AI systems recommendations.	Inappropriate treatment plan.	Patient is exposed to wrong treatment.
Due to the richness of user interface overly complex or information-dense output the user is overwhelmed not recognizing the critical information necessary to make the right medical decision.	User erroneously decides to submit the patient for an additional, but unnecessary CT scan.	Additional x-ray radiation.	Patient is exposed to unnecessary radiation.

Table 9 - Path B - How a user's misinterpretation of the AI system's output can lead to a use error and ultimately a hazardous situation.

5.4 Identify and describe Hazard-Related Use Scenarios

Based on the potential use errors, hazards and hazardous situations identified in the previous steps the IEC 62366-1 standard requires to identify and describe hazard-related use scenarios. Hazard-related use scenarios are essentially narratives illustrating where the interaction between the device, the user, and the environment could lead to hazardous situations and potential harm. These use scenarios help product development teams understanding how devices are used in real-world situations and identifying how potential hazards may arise during the use of medical devices.

Each hazard-related use scenario represents a sequence of events where the interplay between the device, the user, and the environment could lead to a hazardous situation and potential harm. The information analysed so far, such as contributing factors (section 4), use errors (section 5.2), hazards and hazardous situations (see 5.3) will be essentially stitched together to form an illustrative narrative that highlights critical paths leading to use errors and which would ultimately result in a hazardous situation.

The following tables show three concrete examples of hazard-related use scenarios, and a potential way how to construct them. They illustrate how various elements of information analysed so far can be linked together into one scenario. Available information elements are:

- contributing factors (as analysed in section 4)
  - use environment,
  - AI system properties,
  - their impact on perception,
  - their impact on cognition
- resulting use errors (as analysed in 5.2),
- hazards and hazardous situations (as analysed in 5.3):

Element	User Task ➡	AI System Property ➡	Cognition ➡	Use Error ➡	Hazard/Hazardous Situation
Use Scenario	A physician diagnoses a patient.	The user relies on an AI system for patient diagnosis. Due to its **probabilistic nature**, occasionally provides varying outcomes for similar cases.	This variability is not recognized by the user, leading to **overtrust** in the system's accuracy.	Consequently, the user **fails to perform additional diagnostic tests** when the system provides a low probability of a serious condition….	…resulting in a missed critical diagnosis.

Table 10 – Example Use Scenario: Missed Diagnosis Due to Overreliance on AI Probability Assessments

Element	User Task ➡	AI System Property ➡	Cognition ➡	Use Error ➡	Hazard/ Hazardous Situation
Use Scenario	A nurse performs patient assessments in an emergency room.	An AI-based emergency room triage system presents a user with an **overload of information**, …	… significantly increasing **cognitive load**.	The overwhelmed user **delays the decision** to escalate care for a patient showing subtle signs of a life-threatening condition, as the AI's prioritization did not flag it as immediate, …	… leading to deterioration in the patient's condition.

Table 11 – Example Use Scenario: Critical Care Delayed by Information Overload in Emergency Triage

Element	User Task →	Cognition →	Cognition →	Use Error →	Hazard/ Hazardous Situation
Use Scenario	A clinician makes recommendations for a personalized treatment plan	The clinician uses an AI tool for recommending personalized treatment plans. The tool's output, based on complex data analysis, does **not align with the clinician's mental model**, built from years of clinical experience.	This mismatch leads to the **misinterpretation** of the recommended treatment as being less effective, …	… causing the clinician to **choose** a more conventional but **suboptimal treatment** pathway for the patient.	The patient receives suboptimal treatment.

Table 12 – Example Use Scenario: Suboptimal Treatment Choice Stemming from AI Recommendation Misinterpretation

By crafting such detailed scenarios, development teams can better understand the practical implications of AI integration into medical devices. This understanding is crucial for the design but also for creating test protocols that accurately reflect real-world interactions, thereby enhancing the relevance and effectiveness of usability testing.

5.5 Select the Hazard-Related Use Scenarios for Summative Evaluation

The next process step is to select hazard-related use scenarios to prioritize which scenarios will be evaluated in detail during the summative evaluation. In the authors view, this process step of selecting critical scenarios based on risk analysis is not fundamentally different for AI devices and therefore no further guidance is provided here. By focusing on high-risk scenarios, manufacturers ensure that the summative evaluation effectively assesses the safe use of the medical device's user interface.

5.6 Establish User Interface Specification

The user interface specification should define clear, testable technical requirements for the user interface. This includes requirements for user interface elements associated with risk controls, focusing on mitigating the contributing factors identified in Section 4 to prevent use errors.

A comprehensive user interface specification should address all hazard-related use scenarios at a minimum. By defining user interface design elements, it specifies how the user interface should support safe and effective interaction within these scenarios, ultimately mitigating the risks associated with use errors.

Furthermore, the user interface specification should indicate whether accompanying documentation and/or user training is required to educate users on safe and effective interaction with the AI system. This documentation and training can help users to understand the AI's capabilities and limitations, as well as how to interpret specific outputs within the context of use. Beyond understanding and correctly interpreting system output, the device's user interface, accompanying documentation and training must be designed to ensure users also understand:

a. that the output is based on AI technology,
b. how their input and other data contribute to those output,
c. any explanation provided by the system regarding its decision-making process.

An effective user interface design plays a crucial role in mitigating these risks. The following table provides a list of examples of specific user interface design considerations related to AI-enabled user interfaces. These considerations address potential use errors associated with the contributing factors identified in Section 4 to promote safe and effective interaction within hazard-related use scenarios. The list is not meant to be exhaustive, but rather wants to demonstrate how potential causes for use errors can be translated into user interface features.

Potential Causes for Use Error	**User Interface Design Considerations**
Lack of explainability or algorithm bias leading to misinterpretation of AI outputs, potentially delaying critical treatment decisions.	Interactive Explainability Panels: Provide on-demand explanations of the AI's reasoning process behind a particular output. This could include visualizations of the data used by the AI, rule-based explanations, or counterfactual examples.
Probabilistic nature of AI systems leading to overreliance on AI outputs without considering inherent uncertainty, potentially leading to missed diagnoses.	Probabilistic Output Display with Confidence Intervals: Present AI outputs alongside visualizations of their confidence intervals (e.g., error bars, range estimates). Highlight the probabilistic nature of the outputs and how they should be interpreted alongside other clinical information.
Lack of context information leading to missed or incomplete data entry, potentially causing inaccurate AI outputs and incorrect treatment plans.	Context-Aware Prompts: Integrate context-sensitive prompts within the user interface to guide users for additional information input when the AI system detects a lack of necessary context for accurate outputs. Examples could include prompting users for specific vital signs or historical medical data.
Richness of user interface or information overload leading to inattention to critical information due to overwhelming data displays, potentially resulting in overlooking crucial details for diagnosis or treatment.	User-Customizable Information Filtering: Allow users to customize the level of detail displayed in the user interface based on their needs and preferences. This could involve filtering data visualizations, collapsing or expanding information sections, or prioritizing critical information for specific tasks.
Multi-modal user interface leading to errors due to user limitations with traditional input methods (e.g., visual impairments), potentially causing delays or inaccuracies in data entry.	Multi-Modal Input Options: Offer alternative input methods beyond traditional keyboard and mouse interfaces. This could include voice commands, touch screen interaction, or integration with wearable devices for data input.
Interdependence between context of use and user interface leading to confusion or errors	Adaptive User Workflows:

Potential Causes for Use Error	User Interface Design Considerations
due to mismatch between user expectations and the user interface for a specific task, potentially causing delays or incorrect actions.	Design user interfaces that can adapt to different workflows associated with various AI functionalities. This might involve offering different user interfaces for initial diagnosis, treatment planning, or ongoing patient monitoring.
Algorithm bias or limitations in AI capabilities leading to underuse of the AI system due to inability to address potential biases, potentially resulting in suboptimal treatment decisions.	User Override Mechanisms: Incorporate mechanisms that allow users to override AI suggestions or recommendations when their clinical judgment dictates otherwise. This could involve functionalities like flagging concerns, providing justifications for overrides, or reverting to alternative decision-making pathways.
Complexity of user interface or lack of user experience best practices leading to misinterpretations of information due to unclear visual design, potentially causing delays or incorrect actions.	Clear Visual Design Principles: Employ clear and consistent visual design principles throughout the user interface to avoid misinterpretations or confusion. This includes using intuitive icons, consistent colour coding, and appropriate information hierarchy.
New or changed data entry methods associated with AI interaction leading to data entry errors, potentially causing inaccurate AI outputs, and impacting decision-making.	Automated Data Validation Checks: Integrate automated checks within the user interface to ensure data entered by users conforms to expected formats and ranges. This can help mitigate errors due to typos or incorrect data input.
Lack of awareness of system limitations or implications of continuous learning leading to user to over-trust and over-rely on potentially wrong system output.	Clear Disclaimers and Limitations: Prominently display disclaimers within the user interface that communicate the limitations of the AI system (e.g., potential for errors, dependence on training data). This educates users on the appropriate use of AI outputs and the importance of clinical judgment.
Visual or auditory limitations of the patient leading to misinterpretation of critical instructions or information displayed on the AI-enabled medical device. This could lead to incorrect use of the device.	Accessible Interface Design: Integrate accessibility features to accommodate users with visual or auditory impairments (e.g., screen readers, colour contrast options, audio descriptions). Ensure the user interface complies with relevant accessibility standards, such as ISO 9241-171.
Automation bias leading to overreliance on AI recommendations, potentially resulting in overlooked or incorrect diagnoses.	Sequential Display of AI Results: First, the radiologist completes their initial assessment without AI assistance. The AI system's results are displayed only after the radiologist has documented their findings, providing a second opinion. This approach

Potential Causes for Use Error	User Interface Design Considerations
	encourages independent decision-making and reduces the risk of automation bias.
All contributing factors potentially leading to unidentified use errors or limitations in the initial user interface design, which could compromise patient safety or treatment efficacy.	User Feedback Mechanisms: Integrate functionalities within the user interface for users to provide feedback on the AI's performance, report errors, or suggest improvements. This could be used to identify and address unforeseen use errors over time and improve future iterations of the AI system.

Table 13 - Examples of specific user interface design considerations related to AI-enabled user interfaces.

The design considerations like shown in this table provide a valuable foundation for developing a user interface specification for the AI medical device. However, to comply to clause 5.6 of the IEC 62366-1 standard these considerations still need to be translated into a set of clear, testable requirements. These testable requirements are essential to guide development teams and ensure each element of the UI functions as intended. While the process of transforming design considerations into testable requirements is a vital step, it's important to acknowledge that this process as part of requirements engineering is not specific to AI devices. It applies to the development of user interfaces for all medical devices and therefore falls outside the scope of this discussion, which focuses on considerations unique to AI technology.

5.7 Establish User Interface Evaluation Plan

Clause 5.7 of the IEC 62366-1 standard mandates that manufacturers of medical devices establish and maintain a detailed User Interface Evaluation Plan. This plan ensures the user interface, especially for AI technologies, is safe and effective. It includes documenting objectives, evaluation methods for formative and summative evaluations, and criteria for usability tests, including user group selection and test conditions.

There are two main approaches to User Interface Evaluation:

- **Formative Evaluations:** Iterative assessments during development to refine the user interface. These evaluations involve prototypes and various methods like expert reviews and simulated use testing, aiming to gather feedback for continuous improvement.
- **Summative Evaluations:** Conducted in the final stages of development with production-equivalent devices. These evaluations

Design Validation: Evaluating the User Interface vs. AI Performance

It is crucial to distinguish between user interface evaluation and AI performance evaluation during design validation. User interface evaluations focus on ensuring the usability and safety of the interface for end-users, typically involving usability studies with representative users. In contrast, AI performance evaluations assess the technical effectiveness of the AI system, including data models and algorithms, often requiring real-world users to review AI-generated results.

While it might be logistically efficient to combine these evaluations in a single session, it is essential to conceptually separate and document them distinctly. This distinction ensures clarity in regulatory reviews and helps avoid the common confusion between evaluating user interface usability and AI system performance.

confirm that identified use errors have been effectively mitigated through design and risk control measures.

The unique characteristics of AI-enabled medical devices introduce several challenges to user interface evaluation:

- **Probabilistic Nature of AI Systems**:
  AI system outputs can vary with the same input, affecting test protocol repeatability. This characteristic of AI systems can affect the predictability of test participant's responses. Imagine for example an advisory chat bot for clinicians, where users can ask an artificial expert for clinical advice. In one test, the AI might perfectly present the relevant information. However, in another test run with the same user input, the AI might offer a less clear or incomplete response, requiring the customer to rephrase their question or seek further assistance. This variability in system outputs can make it challenging to assess how effectively users can interact with the AI and achieve their goals. Usability testing for probabilistic AI systems might require additional techniques to account for this variability and ensure a comprehensive evaluation to assess the user's decision-making consistency and potential for misinterpretations.
- **Continuous Learning**:
  Summative evaluations are conducted prior to releasing the product to the market. However, the characteristic of a continuously learning AI system can change significantly during production use. In extreme cases this dynamic could lead to use scenarios not previously anticipated. Thus, continuously learning AI systems raise questions about the sufficiency of a single summative evaluation and might require a re-evaluation strategy to account for changes in device behaviour post market release. FDA expresses the potential need for adopting a total product lifecycle (TPLC) approach to the oversight of AI-enabled medical devices during commercialization (FDA, Total Product Life Cycle for Medical Devices, 2023).
- **Planned Changes to the AI Model**:
  Even without continuous learning, planned changes to the AI model after release might necessitate a new summative evaluation, especially if the UI is significantly affected. This begs the question about the significance of a change to an AI model and its effect to the user interface. While smaller changes to the AI model might not alter the general behaviours of the AI system significantly, others might be regarded as significant enough to require a re-evaluation of the user interface. Defining "significant change" for the user interface and AI model can be complex.
- **Richness of User Interface**:
  AI systems can have complex user interfaces with multi-modal user interaction options. Usability testing can't cover every possible interaction path, potentially leaving some areas untested. Increasing the number of test participants helps, but costs can become prohibitive. An example might be an AI system for radiology that allows doctors to view medical scans from various angles and zoom in on specific areas using voice commands and hand gestures. The sheer number of possible voice commands and hand movements would make it impossible to test every combination during usability studies.
- **Multi-Lingual User Interface:**
  AI-based chatbots are increasingly used in healthcare settings to inform patients or advice clinicians. These chatbots often incorporate complex UIs with features like multi-lingual support. While this allows for wider user accessibility, it presents unique challenges for usability testing. Usability testing should consider not only the technical functionality of the translation features but also the potential

for misunderstandings due to cultural and linguistic variations. Thorough usability testing of a multilingual chatbot would ideally involve recruiting participants who speak each supported language. However, this can be expensive and logistically challenging. Focusing solely on the most common languages might leave potential issues undetected in less frequently used languages.

- **Long-term Changes in User Behaviour**:
  The introduction of AI systems can fundamentally alter work patterns and user interactions over time. Initial user behaviours observed during summative evaluations might change long-term during production use. For example, user might be requested through the manufacturer's instructions for use to double-check the responses of the AI system. Initially the users exhibit a natural tendency to mistrust the AI, following the manufacturer's instructions to double-check. However, with experience, they might become over reliant, falling victim to automation bias, potentially leading to unanticipated use errors.
- **User Training Prior to Test**
  One challenge in summative evaluation of AI-enabled medical device user interfaces lies in balancing the need for user training with the goal of simulating real-world scenarios. While AI-systems might require thorough training and therefore might be a pre-requisite of the summative evaluation. Conducting the test immediately after training can artificially distort user performance data due to the user's recent exposure to the training material. To address this, a strategic time interval must be established between user training and test execution. This interval allows for knowledge retention and decay, ultimately creating a more realistic representation of how users with relevant but not perfect knowledge will interact with the system in a clinical setting.

5.8 Perform User Interface Design, Implementation, and Formative Evaluation

Clause 5.8 of IEC 62366-1 focuses on the user interface design and implementation for medical devices. It suggests the use of usability engineering techniques and methods, including formative evaluations, throughout this phase of the process. A key aspect of this phase is the iterative nature of design and development, with frequent user feedback loops being essential.

5.8.1 Cross-Functional Iterative Development

Conducting formative evaluations early and often benefits not just the user interface design but also the AI model itself. Collaboration between usability engineers and AI engineers is key, particularly during the design phase. Integrating those two disciplines is important to ensure alignment and can benefit both mutually. For example, sharing the same panel of medical experts for evaluating both the AI model and the user interface can be cost-effective.

The following describes the breakdown of a typical iterative design process for AI medical devices:

1. **Building Prototypes:** Develop prototypes showcasing specific aspects of the AI user interface. This allows for early exploration of how users interact with the AI's outputs and controls. Those prototypes can also serve for early AI model evaluation.
2. **Formative Evaluation:** Observe user interactions with the prototype to identify usability issues before final implementation. This is vital for ensuring safe and effective use. Focus on how users perceive, interpret and act on the information provided by the AI, including:

- Understanding AI recommendations
- Trust in AI-generated outputs
- Managing incorrect or unexpected AI behaviours

3. **Iterative Improvement Planning:** Based on user feedback, continuously improve the user interface through an iterative design process.

5.8.2 Usability Engineering Techniques and Methods for AI Medical Devices

As discussed in Section 4, many use problems stem from cognitive challenges users face with AI-enabled medical devices. Here are some helpful usability engineering methods that are particularly well-suited for uncovering cognitive issues in formative evaluations for AI-enabled user interfaces:

- **Cognitive Walkthroughs:** Simulate a user's thought process as they interact with the AI system, focusing on how they interpret explanations and make decisions based on AI outputs. This can reveal potential cognitive burden, or misunderstandings related to the AI's functionality.
- **Think-aloud protocol:** During usability testing sessions, users verbalize their thoughts and thought processes as they interact with the prototype. This allows researchers to observe not only user actions but also their underlying thought patterns and potential areas of confusion, particularly when testing AI-enabled medical devices. This method directly accesses the user's thought process, revealing areas where they might struggle to understand the AI system's outputs or complete tasks.
- **Concurrent Probing:** This technique involves asking probing questions while users interact with the prototype. These questions aim to understand user thought processes, uncover potential misunderstandings, and identify any cognitive overload they might be experiencing.
- **Retrospective Interviews:** After interacting with the prototype, interview users to gather overall impressions and experiences. This is a good time to ask specific questions about any cognitive challenges they faced. These interviews allow users to reflect on their experience and articulate any cognitive challenges encountered while using the prototype.

5.9 Perform Summative Evaluation of the Usability of the User Interface

The last step of the usability engineering process as outlined by IEC 62366-1 is the performance of the summative evaluation of the medical device's user interface. This evaluation in many cases requires methods of simulated-use testing. The test script of the usability test guides the test participants through the hazard-related use scenarios identified earlier. The goal is to demonstrate that any use errors and use difficulties are prevented or managed within acceptable risk levels, promoting safe use.

5.9.1 Evaluating Summative Evaluation Data

During summative evaluation, all use errors and difficulties encountered must be thoroughly investigated to identify root causes. Understanding these causes is crucial for mitigating use errors and refining the device design. As anticipated in Section 4, many use errors in AI devices likely stem from cognitive challenges, such as:

- Cognitive Overload
- Mismatch of Mental Model

- Misinterpretation of AI Output
- Lack of Awareness of Limitations
- Lack of Situational Awareness
- Confirmation Bias
- Trust Issues (Overtrust, Mistrust)

User interview techniques are key to understanding these difficulties and collect sufficient data for thorough root cause analyses. However, techniques like "thinking aloud" or "concurrent probing" can disrupt the user's natural flow of action during simulated use and are not recommended for summative evaluation. Retrospective user interviews are the preferred method. These interviews allow users to reflect on their experience after the simulated scenario, providing valuable insights into their thought processes and potential cognitive challenges encountered while interacting with the AI system.

5.9.2 Residual Risk Evaluation

Following data analysis, the standard mandates a residual risk evaluation related to use. This evaluation should consider how the system might evolve due to continuous learning or planned model changes. Any inherent risk related to the medical device design that cannot be eliminated must be disclosed to users. This is particularly true for those related to the limitations of the AI algorithm and its implications for clinical practice. Residual risks inherent to AI systems might include, but are not limited to:

- Risks related to known limitations or biases of the AI algorithm,
- Risks related to limited accuracy of AI algorithm’s outputs,
- Risks related to variability of outputs based on the probabilistic nature of the algorithm,
- Risks related to continuous learning or other model updates.

Inherent residual risks can be disclosed to users through the accompanying documentation as well as through user training.

5.10 Post-Market Monitoring

The IEC 62366-1 standard being designed as a process development standard does not contain any provisions for post-market usability evaluation, which is particularly critical for AI-enabled medical devices. Due to the novelty of AI technology, these devices require a higher level of oversight to ensure their safety and effectiveness. A single summative evaluation might not be sufficient to address the unforeseen challenges and risks that may emerge as these technologies are deployed in real-world settings. Proactive post-market monitoring is essential to ensure the continued safety and efficacy of the device in real-world use. This monitoring can involve:

- **Collecting and Analysing User Feedback**: Mechanisms should be established to collect feedback from healthcare professionals using the device in clinical settings. This feedback can identify usability issues, unexpected AI behaviour, or potential adverse events not identified during pre-market testing.
- **Monitoring for Performance Degradation**: Monitoring for performance degradation is a critical concern for AI systems, regardless of whether they employ continuous learning methodologies.

Performance degradation can occur when the AI's performance diminishes over time due to changes in real-world data it encounters. These changes can be due to:

- Evolving Standard of Care: Medical practices and best practices can evolve over time. An AI system trained on data from 2015-2020 might become less accurate by 2025 if it doesn't adapt to changes in the standard of care.
- Shifting Workflows: AI systems built on workflows are particularly susceptible to degradation, as clinical workflows and procedures can change over time.
- Long-Term Physiological Changes: Even AI systems based on physiological factors can experience reduced effectiveness. Physiological parameters might change over extended periods, potentially affecting the effectiveness of the AI model.

Techniques for data anomaly detection and performance trend analysis can be employed to identify potential performance degradation in all types of AI systems used in medical devices.

- **Regular Usability Testing**: Supplementing initial summative evaluation with periodic usability testing in real-world settings can be beneficial. This allows for identifying new use errors or difficulties that may emerge as users gain experience with the device and its AI functionalities.

Given these considerations, it is clear that the IEC 62366-1 standard, being focused on development, lacks guidance for post-market usability evaluation. While one approach is to revise the standard to include post-market provisions, another viable option is to explicitly state that the standard does not cover the post-market phase and address these provisions in a separate document.

6 Conclusion

The integration of Artificial Intelligence (AI) and machine learning technologies into medical devices presents a transformative opportunity for healthcare, promising enhanced diagnosis, therapy, and patient care. However, it also introduces unique challenges in usability and human factors engineering, particularly regarding patient safety. This paper demonstrates the effectiveness of the Usability Engineering process outlined in IEC 62366-1 in addressing factors specific to AI-systems which might influence the safe use of the device. It also serves as a guide, emphasizing the specific human factors considerations unique to AI technologies and providing actionable insights for effectively applying the Usability Engineering process.

Transitioning from non-AI to AI systems fundamentally transforms workflows, significantly altering user interactions and responsibilities. This necessitates a tailored approach to usability engineering, accommodating the dynamic nature of AI technologies, such as continuous learning and the probabilistic nature of AI outputs. The selection of applicable usability engineering methods and techniques should be adapted to the specific characteristics of AI technology. For example, contextual inquiry is suitable to gain understanding of the current-state scenarios but might fall short in capturing future-state scenarios essential for AI systems.

The IEC 62366-1 standard's strength lies in its technology-agnostic approach, making it adaptable to AI technologies while necessitating specific considerations for these evolving technologies.

However, as a development-focused standard, IEC 62366-1 naturally does not encompass post-market usability evaluation, which is crucial for all medical devices, particularly AI products due to the novelty of AI technology. One potential solution is to revise the standard to include post-market usability evaluation guidelines.

Alternatively, the standard could explicitly state that it does not cover the post-market phase, with these provisions addressed in a separate document.

By implementing a robust post-market monitoring program, manufacturers can proactively identify and mitigate potential safety issues associated with AI-enabled medical devices, ensuring their continued safe and effective use in clinical practice.

A crucial takeaway from this exploration is the importance of interdisciplinary collaboration. Usability engineers, risk managers, and AI engineers must work synergistically, leveraging expertise from each of their respective fields to navigate the complexities of AI. This collaboration ensures that usability engineering remains focused on mitigating use-related risks and maximizing device safety and effectiveness.

In conclusion, as AI technologies mature and become indispensable in medical devices, the principles of Usability Engineering as defined by IEC 62366-1 offer a robust framework for addressing these novel challenges, albeit with necessary adaptations to address the unique challenges posed by AI technologies.

7 Literature and Further Readings

Baird, P., Nayak, R., & Nienburg, K. (2019). *Building Explainability and Trust for AI in Healthcare.* Xavier Health.

Cooper, A., Reimann, R., Cronin, D., & Noessel, C. (2014). *About Face: The Essentials of Interaction Design* (4th. ed.). Wiley Publishing.

Dr. Sujan, M., Prof. Baber, C., Prof. Salmon, P., Pool, R., & Chozos, N. (n.d.). *Human Factors and Ergonomics in Healthcare AI.* UK: Chartered Institute of Ergonomics & Human Factors.

FDA. (2020, Oct 22). Artificial Intelligence (AI) and Machine Learning (ML) in Medical Devices. U.S. Food and Drug Administration.

FDA. (2021, Jan). Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan. U.S. Food and Drug Administration.

FDA. (2023, Jun 12). *Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices*. Retrieved Apr 20, 2024, from https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices

FDA. (2023, Jun 09). *Total Product Life Cycle for Medical Devices*. Retrieved Apr 20, 2024, from https://www.fda.gov/about-fda/cdrh-transparency/total-product-life-cycle-medical-devices

IEC 62366-1. (2020). *Medical devices - Part 1: Application of usability engineering to medical devices*. IEC.

ISO 14971. (2019). *Medical devices - Application of risk management to medical devices*. ISO.

ISO/IEC 22989. (2022). *Information technology - Artificial intelligence, Artificial intelligence concepts and terminology*. ISO/IEC.

Itri, J. N., & Patel, S. (2018, 3 12). Heuristics and Cognitive Error in Medical Imaging. *American Journal of Roentgenology, Volume 210*(Issue 5). doi:10.2214/AJR.17.18907

Parasuraman, R., & Manzey, D. (2010, Jun). Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors - The Journal of the Human Factors and Ergonomics Society, 52*(3), 381-410. doi:10.1177/0018720810376055

Rajan, J., & Redmill, F. (1996). *Human Factors in Safety-Critical Systems.* Butterworth-Heinemann.

Staggers, N., & Norcio, A. (1993). Mental models: Concepts for human-computer interaction research. *International Journal on Man-Machine Studies, 38*, 587-605.

Tomsett, R., Braines, D., Harborne, D., Preece, A., & Chakraborty, S. (2018). Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems. *ICML Workshop on Human Interpretability in Machine Learning.*

Author's Contact Details

- Michael Engler
  michael.engler@benkana-interfaces.com
  Benkana Interfaces GmbH & Co. KG
- Prof. Dr. Christian Johner
  christian.johner@johner-institut.de
  Johner Institut GmbH
- Martin Krepcke
  martin.krepcke@siemens-heathineers.com
  Siemens Healthineers AG
- Dr.-Ing. Manuel Isaac Martinez Torres
  manuel.martineztorres@philips.com
  Philips Medizin Systeme Böblingen GmbH
- Martin Stangenberg
  martin.stangenberg@stryker.com
  Stryker Leibinger GmbH & Co. KG