Measurements, Algorithms, and Presentations of Reality: Framing Interactions with AI-Enabled Decision Support

NIELS VAN BERKEL, Aalborg University, Denmark
MAURA BELLIO, University College London, United Kingdom
MIKAEL B. SKOV, Aalborg University, Denmark
ANN BLANDFORD, University College London, United Kingdom

Bringing AI technology into clinical practice has proved challenging for system designers and medical professionals alike. The academic literature has, for example, highlighted the dangers of black-box decision-making and biased datasets. Further, end-users' ability to validate a system's performance often disappears following the introduction of AI decision-making. We present the MAP model to understand and describe the three stages through which medical observations are interpreted and handled by AI systems. These stages are Measurement, in which information is gathered and converted into data points that can be stored and processed; Algorithm, in which computational processes transform the collected data; and Presentation, where information is returned to the user for interpretation. For each stage, we highlight possible challenges that need to be overcome to develop Human-Centred AI systems. We illuminate our MAP model through complementary case studies on colonoscopy practice and dementia diagnosis, providing examples of the challenges encountered in real-world settings. By defining Human-AI interaction across these three stages, we untangle some of the inherent complexities in designing AI technology for clinical decision-making, and aim to overcome misalignment between medical end-users and AI researchers and developers.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**; **HCI theory, concepts and models**; • **Applied computing → Health care information systems**; **Health informatics**.

Additional Key Words and Phrases: Human-Centred AI, measurement, algorithms, presentation, MAP, MAP model, case studies, healthcare, medicine, decision-making, cooperative AI, Human-AI

ACM Reference Format:
Niels van Berkel, Maura Bellio, Mikael B. Skov, and Ann Blandford. 2023. Measurements, Algorithms, and Presentations of Reality: Framing Interactions with AI-Enabled Decision Support. *ACM Trans. Comput.-Hum. Interact.* 1, 1, Article 1 (January 2023), 32 pages. https://doi.org/10.1145/3571815

1 INTRODUCTION

The adoption of Artificial Intelligence (AI) systems in the medical domain [70], most commonly through the application of Deep Learning or Machine Learning, is typically accompanied by the promise of increased efficiency [122] and a reduction in medical errors [6]. However, this promise is often not realised in clinical practice [65]. Examples include

Authors' addresses: Niels van Berkel, nielsvanberkel@cs.aau.dk, Aalborg University, Aalborg, Denmark; Maura Bellio, maura.bellio.16@ucl.ac.uk, University College London, London, United Kingdom; Mikael B. Skov, dubois@cs.aau.dk, Aalborg University, Aalborg, Denmark; Ann Blandford, a.blandford@ucl.ac.uk, University College London, London, United Kingdom.

Manuscript submitted to ACM

reduced effectiveness when confronted with real-world constraints [54], (racial) biases [81], low levels of patient trust [120], and adversarial attacks that purposefully trigger incorrect medical decisions [32].

These challenges in realising AI-enabled systems, particularly within the medical context, highlight the need to better understand the relationship between real-world phenomena and the interpretation and output provided by AI systems. To support HCI researchers and practitioners, we present the MAP (Measurement, Algorithm, Presentation) model for conceptualising Human-AI interaction based on our experiences in designing and deploying AI systems and the growing literature on this topic. Through the MAP model, we contrast the interpretation of phenomena in the real world as typically carried out by medical personnel with the necessary steps for an AI system to reach an interpretation and present it to clinicians in actionable form. Through this comparison, we highlight user-centred challenges that may arise when relying on AI support systems. The MAP model consists of three steps taken by an AI support system to provide support to end users: *Measurement*, in which a real-world phenomenon is described and recorded by means of one or more sets of data, *Algorithm*, through which the collected data is transformed into a useful outcome signal, and *Presentation*, in which the outcome is presented to the user. For each stage, we highlight challenges faced in designing and deploying AI systems.

Human-Computer Interaction (HCI) researchers and practitioners should play a major role in transforming theoretical AI solutions to successful real world deployments. Following decades in which the fields of HCI and AI barely interacted [12] or competed for funding [40], the technological maturation of AI-based systems means HCI researchers can now study and evaluate these systems in the real world. It has long been recognised that systems that are not perceived as usable will be abandoned by their intended users [43, 118], with uptake further restricted by wariness of AI-based technology among medical staff [70]. As such, HCI researchers and practitioners face the challenges of ensuring that AI systems are usable, fit for purpose and relevant to real-world tasks, and that the needs of the intended end-users are taken into account from the earliest stage of a system's development. However, designing for real-world interaction with an AI system gives rise to novel design questions [28]. For example, system behaviour evolves over time [39]; it is thus challenging to integrate these systems within existing clinical settings [54, 119]. Further, AI systems may provide decision support proactively without waiting for explicit user input [99]. These challenges are further stressed by the health context, in which errors are potentially grave and raise questions of accountability [57] – increasing the need for explainability of AI-based recommendations or behaviour [46, 94].

The remainder of the article is organised as follows. Section 2 presents an overview of related work in Human-AI interaction, focusing on challenges in designing for Human-AI interaction and the integration of AI in clinical contexts, as well as previous frameworks used to describe the processing of information. We introduce the MAP model in Section 3, outlining the three stages of Measurement, Algorithm, and Presentation and highlighting challenges faced in deploying AI in real-world clinical contexts. We illustrate the model's use through cases studies in colonoscopy practice (Section 4.1) and dementia diagnosis (Section 4.2). Our article ends with a discussion on the proposed model and its relation to HCI (Section 5). Specifically, we provide high-level takeaways for the challenges identified in designing, deploying, and evaluating AI-enabled decision support systems. Further, we discuss considerations on user trust towards AI-enabled support systems. Finally, we highlight limitations and opportunities for future work.

2 RELATED WORK

The increasing capabilities of AI technology have given rise to work within HCI on designing and deploying AI-enabled systems. Of particular relevance to this article are studies focused on integrating AI-enabled systems in clinical contexts. Healthcare is often highlighted as a domain in which AI can make a significant impact [53], in part due to the pressure

to reduce costs and improve medical outcomes. Simultaneously, the field faces challenges in terms of AI integration due to, *inter alia*, diverging stakeholder needs, oversimplifications of real-world healthcare processes, and the potentially severe impact of (algorithmic) decision-making. Lastly, we discuss prior work on the framing of information processing.

2.1 Designing for Human-AI Interaction

With the increasing reliance on AI, the HCI community has been working towards a better understanding of designing AI-driven systems. Various papers have contrasted these systems with more 'traditional' digital systems – highlighting that a change in perspective is required. For example, Amershi et al. note that "*AI-infused systems can violate established usability guidelines of traditional user interface design*" [1]. Similarly, Benjamin et al. highlight that "*design research struggles to engage ML [Machine Learning] as a material for design, as the probabilistic inference of models from data patterns withdraws from being present-at-hand*" [11]. Over 20 years ago, Blandford highlighted areas where there needed to be greater convergence between AI and HCI to reason about people's interactions with AI systems [12], including ensuring conceptual fit (or common ground) between people and systems. While the dominant paradigms in AI since 2001 have shifted from rule-based AI working with minimal data to neural network-based approaches that learn from vast datasets, the need for common ground remains. This is illustrated by Smith & Koppel, who analyse the data structure of a widely used health IT system and highlight critical discrepancies between the measures that matter to clinicians and those recorded within the system [96]. The latter are a poor surrogate for the realities of clinical practice. Specific factors that these authors [1, 11, 12, 96] point to are the unpredictability of AI systems, the automated filtering and presentation of information that may result in unintended information loss, and inconsistency over time in response to consistent input patterns. We discuss relevant work on AI explainability, frameworks for understanding and mapping Human-AI interaction, and guidelines that aim to support in designing Human-AI interactions.

In the context of clinical decision support, the explainability of recommendations has been recognised as a critical element in supporting Human-AI interaction. Decision support systems can cover various interventions, ranging from alerts and reminders to recommending clinical pathways [47]. In 2012, before the recent widespread interest in AI systems, Horsky et al. published a review of best interface design practices for clinical decision support systems [47]. This review highlighted the need to avoid a black box approach (*i.e.*, systems in which the internal workings are opaque and only its input and output can be observed), pointing to the goal of cultivating trust within the system user. To do so, Horsky et al. recommend that "*an explanation of medical logic, including formulas for calculating values, should be accessible on demand so that the justification for alerting is transparent and verifiable*" [47]. Similar recommendations are made in the context of AI systems; for example, Yang et al. investigate how clinicians make decisions in the context of a heart pump implant, identifying a perceived lack of need for and trust in clinical decision support tools [119]. Further, they highlight an opportunity for computational support outside the direct decision-making moment. In a subsequent paper, Yang et al. seek to address the critique of clinical decision support tools providing a poor contextual fit [118], proposing what they term an 'unremarkable' form of AI support. This augments existing user routines, in this case by embedding prognostics in the slides of clinicians' decision meetings, rather than introducing a technology that seeks to replace existing work practices.

To aid practitioners in the design of Human-AI systems, several sets of guidelines have been developed. Amershi et al. proposed 18 design guidelines for Human-AI interaction based on input from designers who evaluated the guidelines against existing AI-infused products [1]. Similarly, Liao et al. worked with design practitioners to develop guidelines for explainable AI systems [66]. As stated by Amershi et al., guidelines are typically a generalisation and might not address all types of AI systems. For example, Van Berkel et al. expand on these guidelines by highlighting

user needs in continuous AI-support scenarios [99]. Large et al. provide an example of an application domain that requires specific considerations beyond established guidelines [60]. More generally, Dix encourages designers to design for appropriation [27], allowing users to take ownership of a technology in ways that might not have been initially predicted. This can help reinforce users' adaptability to complex AI-enabled systems.

Finally, several researchers have aimed to capture lessons learned into models or frameworks to provide generalisable guidance to other researchers. For example, Wang et al.'s work on user-centred explainable AI [112] builds on existing theory from Decision Science to propose a framework of explanations for AI decisions or suggestions. The proposed conceptual framework describes how human reasoning processes inform explainable AI techniques – and how AI can help overcome common flaws in our reasoning. The work further aligns with Miller, who states that the main reasons for seeking explanations are to facilitate learning and to generalise observations so that a conceptual model can be constructed to predict and control future similar phenomena [73]. Wang et al. further highlight that AI explanations should consider the typical caveats in human decision-making [112], such as biases introduced by the use of heuristics to speed up decision-making (see Dual-Process Model for more information [21]). They subsequently evaluated their framework in an intensive care unit and concluded that explainable AI should support different reasoning techniques. Yang et al. propose a framework that helps identify whether and how designing a given AI system differs from a 'traditional' digital system [117]. Their framework distinguishes between four levels of AI design complexity, ranging from probabilistic systems (level 1) to evolving adaptive systems (level 4). By applying this framework as an analytical tool, the authors hope to overcome the gap between conventional AI research and conventional HCI design research to identify the distinct challenges in Human-AI interaction [117].

These works highlight some of the challenges that arise when designing for Human-AI interaction. Recognising the diversity of these challenges, we provide the MAP model for reasoning about how AI systems gather data from the real world, interpret that data, and present the resulting computation back to users to support their decision-making. While our focus is on clinical decision-making, we believe the MAP model applies beyond clinical contexts.

2.2 Integrating AI in Clinical Contexts

The research community has identified a plethora of opportunities for AI technology to contribute to improved patient outcomes or a more efficient healthcare system. For example, in radiology, AI has proven helpful in increasing efficacy and efficiency in detection, characterisation, and monitoring of diseases [48]. In surgery, AI can offer real-time clinical decision support, as well as assistance during actual operations [42]. In ophthalmology, AI is showing promise in diagnosis and referral for retinal disease [25]. The use of AI technology is not limited to physical health, as highlighted in an extensive review by Thieme et al. on the use of machine learning in mental health [97].

Amidst high expectations of AI [78], case studies have highlighted the practical challenges medical personnel face when interacting with AI in their daily work environment. For example, Cai et al. identified pathologists' need to contrast the images of patients' biopsies with the biopsies of previous patients to confirm or reject their hypotheses [16]. While AI systems were able to surface similar-looking images, their study highlighted that algorithmically similar images might not be medically similar. As such, the researchers developed and evaluated a 'refinement' tool that allows pathologists to redefine the criteria for similarity, *e.g.* a specific region or visual pattern. Cai et al. conclude that through refinement, doctors obtained "*the agency to hypothesis-test and apply their domain knowledge, while simultaneously leveraging the benefits of automation*" [16]. Gaube et al. compare clinicians' behaviour in radiological tasks when presented with advice claimed to originate from either an AI-based support system or experienced radiologists [34]. While the purported source of the advice did not affect clinicians' performance, clinicians with high expertise (*i.e.*,

radiologists) rated the quality of advice significantly lower when it was said to originate from an AI system as opposed to a human expert. This highlights the algorithmic aversion that may exist among clinical experts. Wang et al. present a case study on an AI-enabled clinical decision support system (CDSS) in rural clinics [110]. This CDSS presents an AI diagnosis of patients, and is integrated with the existing electronic health record system. The results highlight a disconnect between the workflow supported by the system and the real world. For example, Wang et al. point to the fact that many patients require a referral or a refill of their medications [110]; such cases are relatively trivial for the clinician, but the AI support system requires extensive information before giving a recommendation. Given the time pressures medical staff work under, the CDSS is left untouched. These examples highlight the practical challenges faced when integrating AI systems in clinical contexts, where real-world constraints such as room availability [54] restrict the potential effectiveness of AI recommendations. To address these obstacles, Li et al. argue for developing a 'delivery science' for AI in healthcare, in which multiple disciplines collaborate to integrate AI solutions into complex healthcare scenarios [65].

In this paper, we present a model for studying and developing AI systems that support human interaction. The critical nature of healthcare work, combined with widely established work procedures, particular domain knowledge, and complex care scenarios result in unique challenges to technology adoption.

2.3 Framing Information Processing

The growing reliance on information systems has prompted the development of frameworks to reason about systems' effectiveness in capturing and providing interpretable information for users. People continuously and almost effortlessly interpret the world around us – with our brains playing a prominent role in filtering out content deemed irrelevant to us in the moment [77].

One widely established model in system analysis aimed at explaining information processing is the Input-Process-Output (IPO) model [4]. The model describes systems as being subject to an external feed as provided by the system's environment (*i.e.*, input); this external feed can represent a variety of entities, such as user input in computing or raw materials in manufacturing. Subsequently, the system performs a predetermined procedure as based on the input given (*i.e.*, process), which results in the final product (*i.e.*, output) – for example a specific application operation [23]. The system has to fully rely on the provided input, which is always a limited representation of a complex reality and will therefore influence the output in return [23]. The IPO model has been applied to a variety of contexts, such as information systems [4], understanding of human communication [38], and decision-making [83].

Other models that find their foundation in cognitive science have tried to account for a more complex representation of the external reality. Soar [64] and ACT-R [2] are cognitive architectures: specialised systems that, given an input, produce an accurate simulation akin to human cognitive processes to produce an output. One example can be that of activating a set of memory resources and procedural action to solve a complex task – *e.g.*, in aviation [15].

The nuances that characterise AI, particularly in the medical field, made us realise that prior models were insufficient to fully interpret this area of application. The uncertainty in both the observation and interpretation of clinical phenomena, the range of stakeholders, and the highly context-dependent interpretation of observations provide unique challenges faced in the deployment of AI-enabled decision support systems. This encouraged us to develop the MAP model.

3 CHALLENGES IN INTERPRETING MEDICAL REALITY IN AI: THE MAP MODEL

This section introduces the MAP model and its three stages: Measurement, Algorithm, and Presentation. Figure 1 provides an overview of these three stages required for an AI system to interpret any input. We highlight for each stage relevant challenges encountered when designing and deploying AI-enabled decision support systems in the real world. The first conceptualisations of our model date back more than 15 years to work which discusses the roles, value, and validation of computational models of cognition [13] and the nature of 'computational thinking' [114]. The need to reason about the validity of cognitive models (or of simulations of other real-world systems) was evident, but it was unclear how to measure cognition meaningfully and compare the output of a cognitive model (such as Soar [64] or ACT-R [2]) with the outcomes of naturalistic studies. Fast forward to the recent advances in AI, and it is evident that we face similar challenges in relation to confidence in the validity of AI models designed to support decision-making (or make decisions) in safety-critical situations such as healthcare. We found that we could draw on the same ideas of considering the inputs to a computational model, the processing of that model, and the outputs.

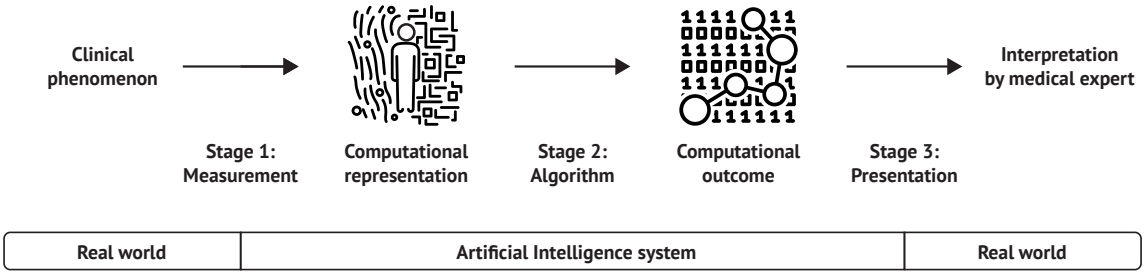

Fig. 1. Systematic representation of the MAP model highlighting the interconnection between the real world and an AI system across the three stages of Measurement, Algorithm, and Presentation.

Information processing systems in medical environments face unique challenges. Interpretation of a patient's status often follows strict procedures to determine the right course of action and offer immediate care when necessary. For example, the 'ABC' protocol for resuscitation prescribes the order in which patients are to be examined (Airway, Breathing, and Circulation) [90]. Through experience and education, medical professionals increase their knowledge on what to look for and which information to filter out. Eye-tracking experiments on medical image analysis have shown that experienced radiologists achieve better results despite having a reduced number of fixations and shorter decision time than their less experienced peers [69]. The ability to directly transfer an observation (*e.g.*, sight, hearing) of a phenomenon into an interpretation allows humans to rapidly build an understanding of their environment, even when faced with an unfamiliar phenomenon.

For AI systems, on the other hand, the interpretation of our surroundings is anything but second nature. AI systems have to be taught what to look for and what judgement to assign to their observations. While trained radiologists will easily spot a broken bone or advanced cancer growth without scanning individual pixels, an AI system requires a computational representation of the image on which it can conduct a computational analysis.

We outline this procedure of computation based on three stages. In the first stage, *Measurement*, analogue information needs to be quantified into measurable data points (by a person or a computer system) for any AI algorithms to be applied to the data. These data points can then be utilised in the second stage, *Algorithm*, in which the AI computes a classification outcome based on a predetermined algorithm and the offered data. In the third and final stage, *Presentation*,

the system provides the user with the outcome of the computation. The interaction does not end there, as the user still has to interpret the presented outcome to determine an appropriate action; this is ultimately the user's goal in using the system.

Conceptually, this has much in common with the IPO model, in that there is an evident mapping between the three stages of each respective model. However, our emphasis is notably different and makes the model less exposed to risky generalisations and misinterpretations. The system receives 'Measurements', indicating that the information entering the system is a limited *representation* of reality rather than an absolute reflection of it. 'Algorithm' plays a similar role to the 'Process' element in the IPO model, though given our focus on AI models and the question of how algorithms are validated, we chose to describe the processing stage as 'Algorithm'. IPO's 'Output' stage focuses solely on how the system sends out the results from the processing stage. In contrast, our focus in 'Presentation' is on the interface between the system and the user. How the system presents or communicates an output that has been generated in machine language to user language is paramount to the user's ability to use that system and information. This includes support for the interpretation of the output and how users can have confidence in the recommendations of AI models and in their interpretation of those outputs to inform clinical practice.

We next detail each stage and highlight relevant challenges that must be overcome to develop Human-Centred AI systems. Figure 2 provides an overview of the challenges for each stage.

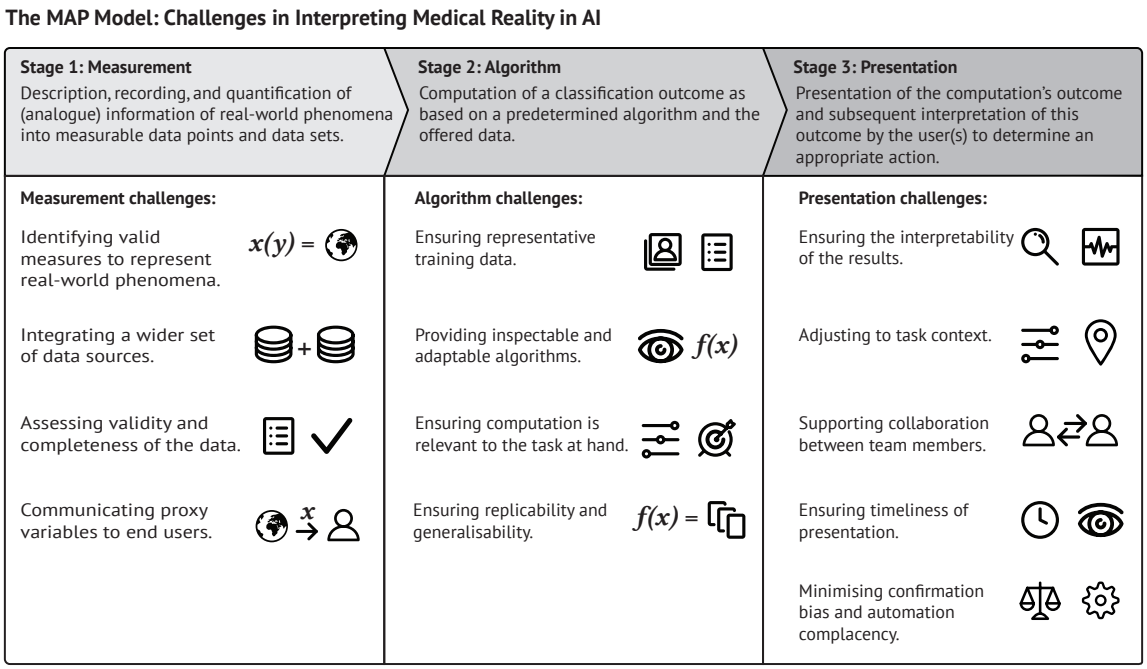

Fig. 2. Overview of the three stages of the MAP model and identified challenges for each stage.[1]

[1]Images based on icons by Shmidt Sergey.

3.1 Stage 1: Measurement

The first step in our model deals with data and measurement, and any inference that requires data. For a healthcare professional, the 'collection' of data can be as straightforward as the observation of a patient (though such data is not necessarily quantitative or computable). Klein refers to this near-instant observation and decision-making, in which individuals rely on their skills and experience, as naturalistic decision-making [58]. Computational systems rely on the conversion of analogue information into digital data. An example of this can be found in the use of heart rate monitors. In the case of ECGs, a measurement of the heart's electrical activity is converted into a digital signal that a trained operator can interpret. In the case of PPGs (photoplethysmogram, as found in finger pulse measurement devices), a measurement of the changes in the light absorption of the skin, caused by the blood being pumped through the body, is converted into a heart rate signal. As illustrated in these examples, neither of which directly measures the heart rate, a computational representation is typically not an exact virtual representation of an analogue phenomenon. Instead, it offers a gateway to obtain the desired information.

This disconnect between reality and eventual representation through measurement, while required by computational systems, raises several challenges for developing Human-Centred AI systems. Based on prior work in this domain, we identified four challenges related to the measurement of analogue phenomena in computational systems.

Challenge: Identifying valid measures to represent real-world phenomena. One of the critical challenges faced in developing and deploying AI systems is the selection of appropriate computational representations. Easily capturable variables might not serve as an adequate proxy for the real-world phenomenon due to these variables' incomplete or incorrect insights. Well-known examples outside the medical context point to a mismatch between conveniently available and seemingly effective proxy variables and the real-world phenomenon of interest. For example, time using an app may be taken as a proxy for engagement, though in other circumstances time taken to complete a task may be interpreted as a measure of how difficult it is. Unsurprisingly, such examples also exist within healthcare, where they can result in severe consequences. For example, Obermeyer et al. describe a widely used commercial algorithm that guides health decisions in the US based on patients' health costs as a proxy for their health needs [81]. However, due to unequal access to care, more money is spent on White patients than on Black patients who are at equal levels of sickness. As such, the risk score calculated by the algorithm provides an incorrect assessment – with Black patients being considerably sicker than White patients at the same score. The system's designers failed to assess the validity of the chosen measure (spend care costs) to represent care needs.

Challenge: Integrating a wider set of data sources. AI-enabled systems frequently rely on the integration of multiple, extensive data sources, often referred to as *big data*. This enables clinicians to identify trends that are typically not detectable without AI support, generating better and more quantitative insights into a particular condition, creating greater consistency and the potential for systematic improvement over time. The combination of data on prior medical outcomes with a detailed patient profile can ultimately support the development of personalised medicine [26]. Despite these prospects, integrating a variety of data sources can be hard to manage clinically [24] and can introduce unexpected and unintended confounds. Similarly, data collected as part of a clinical trial may not be directly comparable to real-world data due to differences in the data collection protocols.

Challenge: Assessing validity and completeness of the data. Health data can be notoriously inaccurate and incomplete. As an example of the inaccuracy of healthcare data, Maling et al. analysed the 2011–2020 records of the UK's National Hip Fracture Database (NHFD) for errors [67]. Over 20% of joint replacement records (across a total of 4531 records) were found to be incorrect, with the most common error found in reporting of the cementing technique.

In addition, missing data can be common. In an analysis across four self-report studies, a widely used method for collecting long-term insights into an individual's well-being, Van Berkel et al. found a substantial difference in the number of responses collected between participants [101]. Across all studies, 50% of the total data was collected from roughly 30% of respondents. Invalid or incomplete data used to train AI systems can cause serious harm. As highlighted by Maling et al. in their analysis of the NHFD, invalid data "*has serious consequences for evidence based on 'historic' NHFD data, causing ambiguity in the ongoing debates over choice of surgery, implant or cement for hip fractures*" [67]. Further, invalid or incomplete patient data used as input to an AI system can result in inappropriate and potentially harmful recommendations.

Challenge: Communicating proxy variables to end users. Finally, the variables used as computational representations of real-world phenomena are often unknown by the end users. This hampers people's ability to construct a mental model of a system's behaviour, which has been highlighted as critical to supporting the interaction with a system [17, 96]. Smith & Koppel describe how healthcare information technology is "*both a medium of communication and a representation of much information—some of which is conflicting, some of which is missing, and all of which interacts with the mental models of designers and users. It is both a microcosm of medical care and it shapes medical care.*" [96]. Providing end users with an understanding of the variables used can prevent users from constructing an incorrect mental model of the AI system's behaviour and having misconceptions regarding the inner workings of AI systems [115]. The transparency of AI systems is receiving increased legislative attention, and is highlighted in a recent proposal for regulation by the European Commission [31].

3.2 Stage 2: Algorithm

The second stage of our model deals with algorithms, with day-to-day healthcare relying heavily on both in-the-moment and long-term reflective algorithmic computation. For example, to assess a patient's pulse, a healthcare professional typically calculates the patient's heartbeat per minute based on observation. Then, through an established categorisation 'algorithm', the heart rate is assessed together with the patient's age and sex to identify whether the heart rate is concerning or within an acceptable range [124]. Best practices for care and observation are determined based on historical successes and failures. This practice is well established, with one of the most well-known examples found in the pioneering work of Florence Nightingale. She collected, interpreted, and visualised data on sanitary practices to propose actionable recommendations to improve health standards in the ward [71]. AI systems similarly rely on historical data to interpret and assess any new information that is presented.

In brief, the development and testing (or validation) of an AI algorithm for deployment in clinical practice typically involves several stages 'from bench to bedside', with iterations to update the algorithm: initial training of an algorithm using a training dataset; technical validation using an independent test dataset (often drawn from the same larger dataset as the training data); and external (or clinical) validation in a chosen clinical context, using locally available data [49]. The algorithm may be further refined as it is deployed in clinical practice, applied to the data for a particular patient. As widely discussed in the AI and HCI literature, reliance on past datasets and inadequate validation processes raises several concerns.

Challenge: Ensuring representative training data. The outcome of any algorithm relies on the data on which it is trained, and is, therefore, inevitably a reflection of what this data represents [41, 50]. As highlighted by Caruana et al., historical healthcare data can be highly misleading [18]. Caruana et al. discovered that a rule-based system learned that patients with pneumonia that also have asthma have a lower risk of dying than the general population. This surprising and incorrect prediction was based on historical events in which patients with asthma and pneumonia would receive

immediate care; this was so effective that the mortality rate dropped below that of the general population. However, if such a system were deployed in the real world, patients would be erroneously turned away from the hospital based on low predicted risk. Another common scenario refers to the datasets used for face detection and tracking. Often, training data excludes marginalised groups, as reported by Buolamwini & Gebru in their report on accuracy disparities in gender classification software [14].

Challenge: Providing inspectable and adaptable algorithms. Healthcare settings require, perhaps more than almost any other AI application domain, that models' outputs are verifiable by clinical experts. Cai et al. state that AI systems are unlikely to perform without error, and that users would therefore benefit from the ability to modify and steer the algorithm in its computation [16]. Hu et al. note that data from a new source (*e.g.*, different hospital) may have different properties from earlier datasets [49]. Despite the ongoing advances in AI technology, this situation will likely remain for some time. Similarly, users may require different support from the AI system as their task progresses, such as moving from a search task to an explanation task. By adjusting the support provided by AI systems, users optimise the tool to fit their current needs and provide new data points that can be used to optimise the support system for future encounters.

Challenge: Ensuring computation is relevant to the task at hand. Despite the promise of AI applications, clinicians are often concerned about their practical application. As Bellio et al. note, "*nominally significant results do not necessarily constitute clinically meaningful or informative findings at the population, group, or individual level*" [10]. Yang et al. conducted a field study exploring clinicians' current decision-making practices related to heart pump implantation and the potential for a CDSS to support these decisions [119]. One of the most significant barriers for adoption was a mismatch between the algorithm's quantitative prediction and the clinicians' information need, which is mostly focused on critical factors that are not captured by the system. In the analysed case, the CDSS was most helpful in automating the retrieval of patient data prior to a clinical meeting rather than actual decision support.

Challenge: Ensuring replicability and generalisability. Replicability and generalisability are amongst the essential requirements for the widespread adoption of AI-enabled decision support systems. Replicability is needed to verify that the outcomes from an AI application in a healthcare scenario remain consistent even after numerous repetitions of the same procedure [108]. Vollmer et al. acknowledged the relevance of this factor in their TREE model, a checklist of 20 critical questions to address transparency, replicability, ethics, and effectiveness [108]. Generalisability, on the other hand, promotes the scaling up of a system developed not only for a specific condition or population but also allows for broader applicability [87]. Predictive modelling, for example, struggles to meet scalability criteria, as datasets used for training and validation are often customised and processed [37], limiting their shift to a different domain. For example, the system developed by De Fauw et al. could make a reliable diagnostic prediction when using Optical Coherence Tomography (OCT) scans from one machine, but not from another, implying that a normalisation process was needed on each machine to ensure generalisability [25]. In their work, Yang et al. investigated the most suitable characteristics for a CDSS in the heart surgery field [118]. However, in defining their key design requirements, they explored other medical domains to ensure the tool's scalability.

3.3 Stage 3: Presentation

The final stage of our model describes the presentation of information to the user, through which the user interprets the results to infer the real-world meaning. This makes the presentation of information critical in deploying AI support systems. While the HCI literature has recently provided a large number of studies and insights regarding the display, explainability, and interpretability of the outcomes of AI systems, only a fraction of this work has focused on deployment

in a healthcare context (see Section 2.2 for a discussion of notable exceptions). Consequently, how to embed AI support in the often hectic day-to-day workflow of medical experts remains an open research area.

Challenge: Ensuring the interpretability of the results. In the medical context, the interpretability of results is often critical. Visualisations can play an important role in AI interpretability, for example when highlighting areas that 'tipped off' the AI in medical imaging. However, supporting end-user interpretation can take many forms, including, for example, text-based explanations [102], interactive applications to contrast AI behaviour across different configurations [113], or through conversation with a conversational agent [51]. Here, we consider any information used to clarify the behaviour of an AI system to the user as contributing to the interpretability of AI systems. Given the often critical nature of clinical decision-making, intelligible models (which can be more easily explained) are favoured or even required over more accurate but non-intelligible models (*e.g.*, random forests, neural nets) [18]. Existing tools on interpretability are often aimed at the designers of the algorithms (see *e.g.* Google's What-if-Tool [113] or Microsoft's InterpretML [79]) rather than the end-users of these systems (*i.e.*, clinical staff). For clinical end-users, AI recommendations need to not only be interpretable, but also lead to clear implications for (medical) action.

Challenge: Adjusting to task context. The situations in which AI systems are deployed within the medical context are diverse. Even within a single procedure, clinicians typically perform several tasks. Consequently, it is critical for AI systems to present the right information at the right time. This includes prioritising the presentation of warnings or alerts but also intelligently switching between different types of support that the clinical team might require.

While existing recommendations on Human-AI interaction have long stressed the need for AI suggestions to be accompanied by explanations (see *e.g.*, Amershi et al. [1]), such explanations can also be experienced as interrupting or even harmful when the user is currently mentally and physically occupied, for example during surgery [103].

Challenge: Supporting collaboration between team members. Existing research on supporting collaboration between team members when engaging with an AI system is limited. Guidelines on designing for Human-AI interaction, such as those presented by Amershi et al. [1], fail to account for teamwork collaboration—focusing instead on individual user interaction with AI systems. Given the collaborative nature of medical practice, AI systems need to support the coordination of tasks between individuals [88]. Malone & Crowston highlight the importance of coordination in designing cooperative work systems [68]. They subsequently define coordination as "*the act of managing interdependencies between activities performed to achieve a goal*" [68]. How to manage these interdependencies, which can necessitate prerequisite activity, demand shared resources, or require simultaneity of activity, in the context of medical AI systems is an open research question.

Challenge: Ensuring timeliness of presentation. AI calculation is not always instantaneous, either due to hardware constraints on location, the need to transfer information back and forth between locations, or simply the complexity of the analysis required. This can introduce challenges, as clinicians may need to decide promptly. Such instantaneous decision-making might occur during surgery but could also include other contexts in which urgency is required due to the patient's or colleagues' availability. While relevant to all contexts, Wahl et al. highlight how resource-constrained settings, in particular, may face challenges concerning the timeliness of presentation [109]. Wahl et al. point to the reliance on hand-written healthcare records, limited access to the internet, and varying connection stability as some of the challenges that may inhibit timeliness in AI support. It is important to note that the timeliness of presentation requires the availability and access to the correct data (Measurement) and a sufficiently quick analysis of these data (Algorithm).

Challenge: Minimising confirmation bias and automation complacency. Clinicians may rely too much on the CDSS's advice without verifying the opposite statement, especially when the provided recommendation aligns

with their expectations [19]. This phenomenon is called *confirmation bias*. One example is provided in a study by Lehman et al., in which they compared breast cancer diagnosis by a group of radiologists with and without CDSS availability [63]. The use of the system led to a higher rate of false negatives than diagnoses without CDSS. A related concept is *automation complacency*, which describes the tendency of people to neglect errors in systems that have generally been reliable [19].

4 CASE STUDIES

How does our model apply in a clinical setting? As well as referencing existing literature (above), we present two case studies, both of which involved exploratory AI systems aimed at experts in the healthcare sector. The first case study, which focuses on colonoscopy practice, is part of a project that aims to support endoscopists in identifying polyps in the colon during an inspection [99, 103]. These polyps are often challenging to identify, with up to 27% of polyps left unidentified [105, 125]. The second case study focuses on dementia diagnosis and interpreting patient biomarkers to predict a patient's neurological decline and identify a suitable care trajectory [8, 9]. Both case studies cover all three of the presented model stages and focus on individual clinical encounters.

4.1 Case study I: Colonoscopy

Collaborating with endoscopists and an AI startup company, we studied AI-enabled prototype applications for colonoscopy. During a colonoscopy, a clinician inserts a flexible tube inside the body. This tube, known as an endoscope, contains a camera, light, and track to deploy various tools (*e.g.*, a snare) or run water through to flush the colon. The procedure typically involves the clinician inserting the endoscope to the beginning of the colon, known as the cecum, after which the clinician slowly retracts the endoscope and carefully inspects the colon wall for possible polyps.

We designed and evaluated different visual markers to highlight areas of interest as detected by AI systems [99]. We overlaid these different visual markers on recorded real-world patient footage, resulting in realistic videos of an AI support system in a colonoscopy context. We subsequently deployed these videos in an online survey targeting relevant clinical staff (N = 36) to evaluate these designs and gain insights into the way these clinical staff envision AI support. Our analysis focused on participants' self-reported ability to detect and localise polyps, as well as perceived interference with their task, all of which were reported through Likert-style questions. Finally, participants ranked the various visual designs and commented on the integration of AI support systems into their daily work environment.

In a second study, we assessed the impact of AI recommendations on clinicians during medical procedures [103]. Here, we presented endoscopists with pre-recorded endoscopic videos, which we overlaid with manually annotated recommendations to mimic the behaviour of an AI system. Using a controller, participants could navigate through these videos and provide a clinical assessment of the presented patient footage (polyp or non-polyp, as all videos contained an AI recommendation: true positive or false positive). We captured the navigation behaviour and polyp assessment of 21 endoscopists as they navigated through these videos. Our results highlight that the time between the presentation of the AI recommendation and the participant's clinical assessment is significantly longer for incorrect recommendations. Further, the type of content (*e.g.*, mucus, polyp) significantly affected this decision time. These results highlight the potentially disruptive nature of AI recommendations, which over time can build up end-user frustration and lead to eventual discarding of the AI support tool. This reinforces the importance of reducing false positives in clinical AI recommendation systems to facilitate real-world adoption.

These studies aimed to obtain insights into how clinicians envision integrating AI-enabled systems into their daily practice. While our evaluations did not involve patients, given clear medical and ethical concerns, we aimed to ensure a

high level of ecological validity to evaluate how data would be measured and results presented (and interpreted) within live clinical practice [100]. We reflect on our design and evaluation of this case study in relation to the MAP model.

4.1.1 Measurement. During a colonoscopy, an endoscopist's primary task is to identify and locate polyps. The clinician typically removes any potential polyp for further pathological analysis. Failing to identify a polyp can have severe consequences, with early removal of precancerous lesions and early detection of cancers increasing patient survival rates [86]. AI support in this context is, therefore, typically aimed at supporting endoscopists in identifying polyps. To achieve this, AI systems are trained on recorded patient image material. These image data banks are manually annotated and subsequently labelled by clinical experts as either non-polyp or polyp.

Assessing validity and completeness of the data. These AI-enabled systems can only identify polyps when they are present in the view of the endoscope's camera. Highlighting potential polyps in the endoscope's view provides only a partial solution for the problem of polyp miss rates. In addition to the situation where the clinician is unable to reach the starting point of the colon (known as cecal intubation) [35], the clinician may fail in mapping the entire colon. This can be because colonic folds conceal part of the colon wall, because parts are covered in mucus, or simply because difficult-to-reach regions were insufficiently brought into view. AI-enabled solutions that aim solely to identify polyps that are already visible and at the surface of the colon may, therefore, not provide a complete measure of real-world phenomena (hidden polyps). Instead, representing the real-world challenges would require an AI-enabled system to also provide support in guiding clinicians in ensuring complete coverage of the colon.

4.1.2 Algorithm. Polyps are manually demarcated to cover the entire polyp area across all image frames in which the polyp is visible to ensure that a future algorithm can recognise polyps from multiple angles. Additional ground truth information regarding the potential polyp and its histological classification can be obtained only if the area has been removed from the patient's colon and histologically examined. Being a highly laborious and time-intensive task, the open sharing of annotated video material is of high value to the future development of AI support systems. One example of such a dataset is the Kvasir-SEG dataset [52].

Ensuring representative training data. Our clinical collaborators raised concerns about the possibility and consequences of false positives once AI-enabled decision support systems are integrated into clinical practices. Currently available systems frequently show false positives when presented with non-polyp material present in the colon, such as bubbles, stool, or wrinkles on the surface of the colon. Given the frequent occurrence of these features, the system can become highly distracting to the operator—potentially resulting in the clinician ignoring the system's correct recommendations or turning the system off entirely. These false positives result from incomplete training datasets, which render the AI system unable to distinguish between anomalies present in the colon and actual polyps.

Providing inspectable and adaptable algorithms. Following the identification and localisation of a polyp, a subsequent step for the clinician is to classify the polyp's type. Highlighting the need for AI systems to adapt to different end-user sub-tasks, this on-the-spot polyp classification increasingly determines the colonoscopist's subsequent actions. Relatively new practices are that of 'resect and discard', in which diminutive polyps (smaller than 5 mm and almost universally benign) are resected but do not pass through a pathological exam [55], and 'diagnose and leave', in which hyperplastic polyps are left untouched. This may increase patient safety, as any interference with the colon can potentially result in a colon perforation or other unintended side effects [55]. Given the importance of correctly classifying polyps during colonoscopy, AI support is proposed to aid in polyp classification [76]. A relevant HCI question is how to enable endoscopists in switching between the required modes of support (detection vs classification) while simultaneously ensuring that they are not inspecting the colon with the incorrect support system active. During

our interviews, endoscopists expressed concerns that disabling the AI detection support could result in the clinician forgetting to turn the detection back on: "*I don't want to be able to turn it off by mistake without realising it, but it does need to be quick and easy and reliable to do—I would suggest a reminder to turn it on again after a set time (e.g., 1 minute) if it is silenced/switched off*" [99].

Ensuring computation is relevant to the task at hand. The clinicians with whom we collaborated were well aware of the high miss rates reported in the literature [105, 125]. While some clinicians expressed concerns regarding the long-term effects of AI systems on clinicians' capabilities, there was consensus on the need for algorithmic support in identifying polyps. As such, our discussions focused primarily on the successful integration and presentation of algorithmic outcomes.

4.1.3 Presentation. A key characteristic of the endoscopy case is the need for actionable and prompt decision-making since the procedure is not easily paused and later resumed. This need for immediate decision-making stands in contrast to most of the existing work on medical imaging, in which the patient is not physically present during image analysis (see *e.g.* [16]). We propose that this difference in the decision-making procedure requires an alternative perspective on AI support, called continuous AI-support [99, 104], *i.e.* "*systems that 'listen' to a stream of uninterrupted user input rather than individual instructions and can respond to this input throughout the duration of the interaction.*" [104]. In the presented case, the clinician does not proactively request AI support but receives a sustained input of images throughout the procedure, raising novel questions around the presentation of AI assessments.

Adjusting to task context. Alerting the endoscopist to the presence of potential polyps is a key feature of the studied system. It is, therefore, important that the clinician does not miss these alerts. At the same time, the alert should not distract the user from their current task. Through the aforementioned surveying of relevant clinical experts, we evaluated a total of seven unique marker designs (as shown in Figure 3) [99]. The designs of two markers are based on existing systems, namely Figure 3-VI; in which a continuously updating percentage sign indicates the likelihood of a polyp being visible in the current frame, and Figure 3-VII; in which the corners outside the live video change colour to indicate the detection of a polyp. While this display of information in the periphery avoids interfering with the view of the user, the real-world interpretation by the clinician is hampered by missing critical information on the location of the potential polyp. On the other hand, false positives of objects easily identifiable by the human operator as non-polyps (*e.g.*, pills floating in the colon) presented directly on the colon footage might be considered as interrupting. While some of the clinicians reported not being affected by false positives during our evaluation, the consensus was that a large number of false positives would strain clinicians when a system is deployed in the real world. Further, if the user's attention is on the colon, they might fail to notice a change in the corners of the display. As stated by one participant; "*This [design] takes the attention away from the mucosa for a number that is almost continuously changing. [...] This display is very inefficient on its own and with continuous variation*" [99].

Support collaboration between team members. While most research on AI in colonoscopy focuses on supporting the endoscopist in their task of localising polyps, colonoscopy procedures are team efforts that involve multiple medical professionals. Relevant information needs to be presented to all team members. For example, staff nurses and healthcare assistants are responsible for delivering medical instruments through the endoscope when requested by the acting endoscopist, as well as maintaining patient sedation or physically moving the patient [99]. Timely indication to the nursing staff of the needs of both patient and clinician could reduce the procedure time, thereby increasing patient comfort and potentially reducing healthcare costs. Furthermore, the active involvement of experienced nurses as 'second

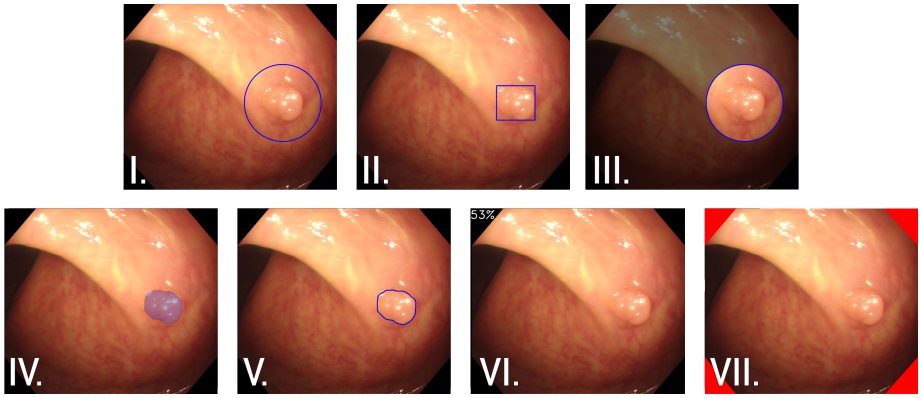

Fig. 3. Visual marker designs to highlight AI recommendations in the context of colonoscopy [99].

observers' during colonoscopy increases polyp detection rates [62]. A critical assessment on supporting collaboration between team members when using an AI-support system in the clinic is currently lacking.

Ensuring timeliness of presentation. The assessment of individual video frames at near real-time speed requires substantial computational resources. Delays in the presentation of the algorithmic assessment are detrimental to the usefulness of the AI system, as clinicians require direct feedback when manipulating the endoscope. Further, a prolonged inspection time would result in increased patient discomfort and reduce the effectiveness of an endoscopy unit. To support the timeliness of presentation in remote and rural areas in which the required computational resources are unavailable, early work has started on using high-speed satellite communication to facilitate real-time AI diagnosis of endoscopic images [98].

Minimising confirmation bias and automation complacency. Assessing the potential impact of increased bias and complacency among medical staff following the introduction of AI support requires a thorough longitudinal evaluation. While this was outside the scope of our studies, we observed that confirmation bias affects colonoscopy differently than binary decision problems in stationary imaging [19, 63]. For example, false positives result in increased clinical effort – while the closer inspection of an area can be considered a positive side-effect, it also leads to higher patient discomfort and healthcare costs. As stated by one clinician: "*Knowing that optical diagnosis is imperfect, I would worry when disregarding an area the AI has flagged as potentially abnormal in case I am wrong. I am a naturally cautious practitioner and worry the extra inspection of potentially normal areas would add a lot of time as I try to really satisfy myself it is the machine and not myself making the incorrect call.*" [103]. Our work focused specifically on the impact of false positives (*i.e.*, an AI suggestion in case of non-polyps) [103]. Clinical outcomes could be severely impacted if clinicians fail to recognise false negatives (*i.e.*, the AI support failing to pick up on polyps that are present) due to increased automation complacency.

Through the presented user studies, we uncovered important implications for the design of AI support in the context of colonoscopy. While it is often challenging to evaluate AI systems in the real world due to ethical concerns, the discussed studies highlight distinct but effective ways for HCI researchers to evaluate different Human-AI interactions. Such outcomes can provide clear-cut takeaways for the design of AI systems (*e.g.*, visual marker, integration into existing practice [99]), as well as novel theory on Human-AI interaction (*e.g.*, designing for continuous interaction [99, 104], dealing with false positives during interaction [103]).

4.2 Case study II: Dementia

The second case study refers to published work on developing a CDSS for clinicians working with dementia patients [7, 8]. Within a European project developing predictive models to understand the course of neurodegenerative conditions, we focused on how one specific model for staging and prediction of dementia progress could be brought to clinical practice.

We adopted a user-centred design approach, starting from a study that focused on understanding clinicians' needs and current practices, their use of and access to data, and their approach to staging and forecasting [8]. This was explored through field observations at multidisciplinary clinical team meetings (where lists of patient cases are discussed and potentially actioned), interviews, and surveys. Interviewees were also exposed to the concept of modelling the progression of dementia to test the perceived value in clinical practice. Our analysis focused on current barriers and facilitators for the clinical adoption of these models.

Based on our initial understanding of the clinical setting and the model's potential, we designed a mock interface for a CDSS and tested the design concept in a second study with specialist neurologists and psychiatrists in dementia care. We aimed to understand their potential use of such a tool, as well as their understanding and construction of a mental model for this innovative and complex concept [7]. We conducted 17 individual sessions with clinicians, in which we asked them to assess a mock patient at baseline and two follow-ups. We observed high agreement from clinicians with the output from the model (74%). Clinicians reported that they would use this tool for various purposes, as discussed in Section 4.2.3.

This set of studies, from the initial problem understanding to prototype testing, added another perspective to our knowledge of AI-enabled systems integration into clinical practice. Whilst clinicians are again the target users for this application, the clinical setting, workflow, requirements, and tasks differ substantially from the colonoscopy case study.

4.2.1 Measurement. The diagnosis of a patient with dementia is a particularly prolonged task [91]. It requires a multidisciplinary team of specialists and a collection of examinations (biological, clinical, imaging), with longitudinal observations increasing the robustness of the assessment. When evaluating a patient for dementia, specialists need to determine the level of cognitive and behavioural deterioration, combined with neurological deterioration. Although guidelines have been defined for diagnosis and treatment [85], dementia diagnosis is heterogeneous, with different centres and specialists using different measures for assessments and monitoring, and with an overall qualitative approach based on experience [61].

Within this context, AI offers several potential improvements. Clinicians would be able to adopt a more quantitative approach to the way data is reviewed and interpreted. Given the high number of indicators needed to consolidate a diagnosis for a given patient, and markers for dementia becoming more sophisticated, AI can combine multiple indicators into an integrated picture of the patient's status and progress [121]. This could allow dementia specialists to identify patterns in the data that would not otherwise be noticeable, promote early detection of the condition, and support the comparison of treatment outcomes [82].

Integrating a wider set of data sources. A unified measurement system for such a heterogeneous condition, however, has its pitfalls. Different centres adopt distinct clinical tests, collect different indicators, or consider dissimilar cut-off ranges for biological markers. Moreover, indicators commonly used in clinical trials, where there is an abundance of clean and complete data for training AI models, might not be available in actual clinical practice [8]. This may result in misrepresentation of how data is thought to be clinically relevant in one context but not in another. Availability of measures in the first place is key to training the model and generating results further down the line. Therefore, chosen

measures for a model should be generalisable or at least adaptable to different contexts and specialisation levels. In the case of dementia, composite scores from clinical tests could be used instead of test scores themselves [10].

In addition, the flow of collecting, analysing, and merging data to be inputted into the model might happen at slightly different times, even though they refer to the same 'assessment window'. For example, there might be delays in receiving a report from radiologists regarding the brain scan, or clinical tests might be performed on different days or times of the day, when the patient status might vary substantially [8]. If these data sources are merged and classified as a single point in time, the picture generated by the model might not be accurate, affecting later stages of decision-making.

Communicating proxy variables to end users. Another source of misrepresentation was the expression 'early stages'. Whilst for the computational scientists 'early stages' indicated the first abnormalities in dementia markers related to biological changes (not clinically observable), for clinicians, the same expression described a patient presenting with initial symptoms affecting cognitive performance. What clinicians consider 'early stages' is more advanced disease than the computational team's view; "*If you put a scale up like that [...] and say right, you've just started having memory problems, it's all very mild, that means you are at stage 5 out of 12, [the patient might reply] 'well, I think you blind me, I am half-way gone!', whereas we know that, although the disease may be present in their brain for 10 years, symptomatically they are at the beginning of that journey.*" [7]. Another aspect of this misrepresentation was evident in one of the early visualisations of the model, with stages represented sequentially from the first (stage 1) to the last (stage 12) marker becoming abnormal. This way of representing stages was not informative to clinicians [8]. In subsequent iterations of the model's visualisation, we replaced the numerical stages with the age of the patient [9]. This way, we could represent the increment in abnormal markers against patient's age without suggesting potentially misleading information on the severity of the condition.

4.2.2 Algorithm. Expert clinicians examining a person with dementia can rely on past experiences that give them the context to diagnose that patient and define a treatment plan. Disease progression models for dementia follow the same principle: they learn how to describe an individual patient's status and possible progression based on a dataset of previously classified patients. Although models might be considered more reliable than human judgement, clinicians reported various concerns when using them for decision-support [8].

Ensuring representative training data. To test a model for any given population, the training dataset needs to be representative of the setting and the reference population. In our case, the training data came from a highly controlled dataset, based on an American observational study [84]. Consequently, applying the algorithm to a different population could result in inaccurate staging and predictions for the patients. Clinicians interviewed in the UK and Belgium asked whether results were validated against their local population [8], and expressed a lack of trust in the algorithm at this stage of the model's development. Although having sufficiently large training datasets for different populations and types of markers for dementia might be unrealistic at present, it is important to document the characteristics of the training data so that clinicians can determine whether or not the dataset is representative for the target population.

Providing inspectable and adaptable algorithms. The Event-Based Model [121] used in our study works under two critical assumptions. The first is that in the training phase, the algorithm generates the most likely sequence of events the patients will go through in their disease progression. An event is characterised by a single disease marker (be it clinical, biological, or imaging-related) becoming abnormal. The second assumption is that the progression is monotonic, meaning that an abnormal marker cannot change back to normal. Whilst a model is an approximation of reality, dementia is a heterogeneous condition, and no patient's journey is like another; "*It can be misleading, I mean, we have to work with some level of uncertainty. And when people ask about progression, we are very cautious in terms of*

saying very strong statements about what's coming in the future, because we know there's some huge variability between patients" [7]. Even if the algorithm's outcome factors in this uncertainty, it should still allow specialists to understand how it came to its conclusions.

Ensuring computation is relevant to the task at hand. When innovation originates from the computational rather than the clinical side, it might be challenging to verify and address the actual clinical need [8], which can result in limited adoption. In this case, the predictive model for staging and forecasting dementia was developed as part of a work-stream building on extensive data collection initiatives [84], but no work had explored the value for day-to-day practice. During a workshop, one of the computational scientists developing the models stated: "*My perspective is that it might be less a question of 'are the models ready?', it's more a question of 'are they needed?'*" [7] We explored the current clinical practice in specialised dementia centres, using field observations and interviews, and displaying an early version of a CDSS clickable prototype [7, 8]. By following this approach, we could frame the gap between the model's output and the clinical expectations, the future intended use in practice, and the barriers to its adoption.

4.2.3 Presentation. In our studies with dementia specialists, we wanted to define the intended use of our model-based CDSS to optimise the *what* and *when* of its inclusion in the clinical workflow. Clinicians reported that they anticipated using the CDSS in three main scenarios: (1) as preparation for upcoming appointments; (2) to support the discussion in multidisciplinary team meetings; and (3) for training.

Ensuring interpretability of the results. To visualise the algorithm within the CDSS (icompass: Fig. 4), we designed the output from the monotonic event-based model as a downward trajectory that decrements depending on the number of markers that turn abnormal. Although this visualisation aimed to pair the characteristics of the model's output with clinicians' familiarity with survival curves, the clinical reality is quite different from this interpretation, with patients' markers fluctuating and following different sequences of events. Nevertheless, specialists reported finding this tool and the visualisation helpful in integrating multiple assessments and aiding individual and collaborative thinking. The model can additionally forecast the future patient trajectory. This is represented with a dashed line for the most likely prediction and an opaque area for the probability range. Whilst specialists are more positive regarding the trajectory to date, which they can validate based on their expertise and observations, the predicted trajectory was more controversial. Clinicians are themselves uncertain of unexpected turns that dementia can take in a patient's life, and they admitted that prediction might be a game-changer when a treatment is available.

Adjusting to task context. Based on clinicians' responses, we designed a user journey that illustrates four levels in current dementia clinical practice (patient interaction, decision-making, data collection/analysis, and peer discussion) and how icompass could be embedded in the workflow. Icompass is also expected to have an impact on speeding up work and easing cognitive workload, thus facilitating adoption in clinical practice; "*It could be useful when it isn't obviously clear from numbers that there is a decline. Just by placing the numerical output of an image registration and of a psychometric registration side by side and giving you some kind of global decline measure*" [7].

Supporting collaboration between team members. Clinicians envisioned a tool that needs to be accessed from different settings and by various team members. Consequently, data displays and descriptions need to be understood by care team members with different roles and expertise levels. Most participants did not expect the system to be used with patients. Hence the system could be used individually by multiple professional figures or as a support tool for multidisciplinary team meetings. As one clinician stated: "*In one MDT session a multitude of cases are discussed and that implies remembering numbers of all tests, eventually compared with previous scores and other information. It is hard to keep track of all information for each*" [7]. By summarising various data into a coherent picture of the patient progress

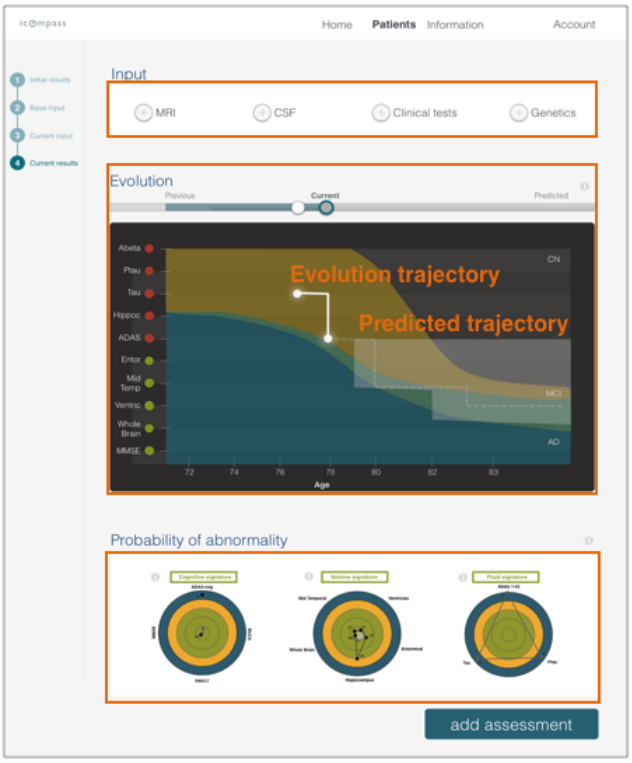

Fig. 4. Example of icompass output screen [7].

and forecast, clinicians could see its value in saving time and supporting the conversation through visualisations of the model.

Based on the findings of the user studies, recognising the current model's limitations, and considering the uncertainty in the prognosis and absence of disease-modifying treatments, we conclude that this system is not yet ready for deployment. Given the facilitated access to more sophisticated markers, controlled data collection, and population access, a clinical trial is a suitable setting to further test icompass, representing a sound transition towards its use in clinical practice [116].

5 DISCUSSION

By introducing AI-enabled systems into the already complex day-to-day operation of medical practice, we risk increasing the complexity of interpreting health-related phenomena and their representation as offered by digital tools. Our MAP model can support high-level reasoning about how AI systems gather data from the real world (including through human data entry), interpret that data, and present the resulting computation back to users to support their decision-making. We have outlined the three stages of the MAP model; Measurement, Algorithm, and Presentation. To support the development of future Human-Centered AI systems, building a clear understanding of these three steps is necessary to ensure adoption and alignment with end-user needs.

Phase	Challenge	Case I	Case II	Selected refs.
Measurement	Identifying valid measures to represent real-world phenomena.			[81]
	Integrating a wider set of data sources.		✔	[8, 24, 26]
	Assessing validity and completeness of the data.	✔		[67, 101]
	Communicating proxy variables to end users.		✔	[17, 96, 115]
Algorithm	Ensuring representative training data.	✔	✔	[14, 18, 45]
	Providing inspectable and adaptable algorithms.	✔	✔	[16, 99]
	Ensuring computation is relevant to the task at hand.	✔	✔	[10, 119]
	Ensuring replicability and generalisability.			[87, 108, 118]
Presentation	Ensuring the interpretability of the results.		✔	[18, 51, 102]
	Adjusting to task context.	✔	✔	[1, 72, 99]
	Supporting collaboration between team members.	✔	✔	[68]
	Ensuring timeliness of presentation.	✔		[109]
	Minimising confirmation bias and automation complacency.	✔		[63, 103, 123]

Table 1. Overview of challenges discussed across the Measurement, Algorithm, and Presentation phases. We indicate their occurrence across the case studies and provide references for additional reading.

Next, we position our work within Human-AI interaction. In Section 5.1, we discuss the application of the MAP model by contrasting our two case studies. Further, we provide concrete takeaways for each of the identified challenges to support in designing, deploying, and evaluating AI-enabled decision support systems. Section 5.2 provides a detailed examination of end-user trust in AI-enabled systems when deployed in real-world clinical systems. Finally, in Section 5.3 we highlight the limitations of our contribution and outline opportunities for future work.

The field of HCI has thus far predominantly focused on the final stage of the MAP model: Presentation. For example, our community has studied how to best design AI-enabled systems [1, 11] and evaluated the interaction with AI and decision support systems through case studies [8, 47, 103, 119]. However, it is critical to highlight that the interpretation of phenomena by AI systems covers all three stages of MAP. While most end-user interaction surfaces at the Presentation stage, a restricted focus on this stage could have a detrimental effect on the design of AI-enabled decision support systems. This is especially true when considering the transition from prototype development and evaluation in the lab toward deployment in the wild. For example, appropriately selected and communicated measurement variables not only ensure the correct operation of the AI system, but can also support end-users in building a correct mental model of the system's operation. Such a model can aid users during regular operations and assist in their reasoning when the system produces unexpected results.

While our article focuses on clinical decision support systems, the MAP model applies to contexts outside the medical domain. The interpretation of real-world phenomena by AI-enabled systems, illustrated in Figure 1, commonly follows the three stages of Measurement, Algorithm, and Presentation. We summarise the challenges introduced and discussed in this article in Table 1, and indicate which apply to each of the two case studies discussed above. Our identification of challenges was made inductively, based on our experiences obtained through the case studies and a review of the literature. As such, not all the challenges apply to both (or even either) of the case studies. Finally, Table 1 also provides pointers to relevant references in relation to the challenges discussed.

Phase	Description	AI-Enabled Decision Support Systems in the Wild
Measurement	Initial stage of an AI system's interpretation of a real-world phenomenon, in which the AI system captures digital measures. Results in a computational representation of the phenomenon.	***Similarities.*** Both cases involve AI-systems designed to align with the existing workflow of the clinicians. ***Differences.*** Clinical observations were carried out on different temporal scales. ***Suggestion.*** Enable users to evaluate the validity and completeness of data and datasets that are input to the AI algorithm.
Algorithm	Transformation of digital representation into a classification as based on predetermined assessment criteria, generally based on historical data. Results in a computational outcome.	***Similarities.*** Both cases are characterised by uncertainties in assessing and diagnosing. ***Differences.*** Homogeneous and locally sourced training data as compared to heterogeneous and internationally sourced training data. ***Suggestion.*** Facilitate algorithm inspection by users through access to pertinent information about training data, validation, and performance.
Presentation	The AI system provides a presentation of the computational outcome, typically intended for the system's end-user to interpret.	***Similarities.*** Both cases augmented the diagnosis by the clinicians; neither case employed automatic AI diagnosis. ***Differences.*** Acute and immediate diagnosis versus a more deliberate and complex diagnosis. ***Suggestion.*** Ensure that the presentation of the output is timely, clinically pertinent, and easily interpreted, including any measures of uncertainty in the outcome.

Table 2. Overview of the MAP model and the questions raised about the use of AI systems in the wild.

5.1 AI-Enabled Decision Support Systems: Overcoming Challenges in the Wild

Based on existing literature as well as our own experience, we identified and illustrated 13 challenges across the three stages of measurement, algorithm, and presentation (Sections 3.1–3.3). Next, we presented the two cases on colonoscopy and dementia as concrete and illustrative examples of 'in the wild' studies of AI-enabled decision support systems. These cases involved user interaction studies with clinicians from distinct health domains. Next, we discuss experiences and considerations for the two cases by highlighting similarities and differences between them. To support researchers and engineers in overcoming these identified challenges in their own research scenarios, we provide high-level takeaways for each stage of the MAP model.

5.1.1 Measurement. The first stage of our MAP model deals with observing real-world phenomena and converting analogue information into digital data. Our cases showed similarities in the way work was conducted with and without AI support, but they also differed significantly on the timescale of measurements. The cases shared similarities in how existing workflows were supported and maintained. Both cases utilised AI decision support to augment current practices and measures rather than drastically transforming practices to align with a new support tool. Furthermore, in both cases, the clinician is still in control of the measurements provided to the AI system: either by directing the endoscope to a specific target area (colonoscopy) or by manually entering data into the system (dementia).

On the other hand, our two case studies employed different ways of observing real-world phenomena and consequently differed in how analogue information should be converted into digital data. They involved different diagnosis procedures, as reflected in the measures used to inform the AI decisions and the integration into the clinical workflow. For colonoscopy, the procedure can be characterised by short, focused interactions with the AI-enabled decision support to inspect patients' colons. Decision-making has to happen immediately and on the spot, with limited opportunity to

revisit prior decisions. In the dementia case, the procedure is more long-term and involves multiple indicators. Thus, for the dementia case, big data and high numbers of variables seem essential for consolidating a diagnosis.

5.1.2 Overcoming Measurement Challenges.

Identifying valid measures to represent real-world phenomena. Appropriate representations can be identified only through a thorough understanding of the application context. This requires engaging deeply with domain experts to establish successful Human-AI interaction. This engagement does, however, introduce some considerations to be made by the researchers or designers. Informants may have various concerns towards the introduction of AI-enabled systems, such as job automation [111], fear of loss of manual skills [103], or legal/responsibility implications [76]. While the behaviour of AI systems might be challenging to envision for stakeholders and designers alike [28, 117], the initial stage of identifying valid measures to represent real-world phenomena is similar to the discovery phase in a user-centred design process. It can, therefore, best be approached through exploratory research methods (*e.g.*, stakeholder interviews, diary studies).

Integrating a wider set of data sources. Researchers and designers may consider how AI systems' measurements can be combined or enriched by qualitative data collected by a domain expert. For clinicians this can, for example, include observations of a patient's current mood or quality of life. Combining purely quantitative outcomes with an expert's contextualisation of these outcomes not only allows for personalisation of the AI outcomes to a given context, but also provides specialists with ownership over the data, potentially fostering adoption. The clinicians involved in our studies often mentioned that the course of a condition is very different for each patient and it is, therefore, inappropriate to reason only around numbers [7].

Assessing validity and completeness of the data. Researchers should consider the validity and completeness of integrated data for the domain. This has been stressed by previous studies [67, 101] and, unsurprisingly, incomplete data constituted a concern for our involved clinicians. Of particular importance within critical domains of Human-AI interaction is to focus not merely on 'regular operations', but pay particular attention to exceptions, edge-cases, and alternative ways of working. Such deviations from the norm are most likely to introduce challenges in AI-enabled support, as they can result in unexpected or undesired outputs. Obtaining a complete (or as complete as possible) overview of work protocols and relevant input data requires a thorough, human-centred understanding of end-users' challenges.

Communicating proxy variables to end users. End user knowledge of what a system is capable of and, equally important, incapable of is crucial to successful Human-AI collaboration [5]. Ensuring that users are aware of and understand the variables used in the operation of AI systems is essential in creating correct mental models. The importance of mental models has been long highlighted in HCI [80]. While existing HCI knowledge on supporting users in creating mental models is therefore of value, Eiband et al. warn that AI systems raise new questions due to their increased complexity; "*if mental models are erroneous, or do not adequately reflect the complexity of a system, users may experience difficulty in predicting and explaining the system behavior*" [29]. While mental models typically develop over time [80], safety-critical AI-enabled systems should enable users to quickly grasp which measures are used in its computation.

5.1.3 Algorithm. The second stage of MAP deals with the algorithm and usually takes historical data to interpret and predict based on new data. Our cases were both characterised by uncertainties in the assessment and diagnosis, while they differed in involving either heterogeneous or homogeneous variables and measures. Unsurprisingly, both cases had uncertainties in the diagnosis and assessment. Ideally, introducing an AI-enabled CDSS would remove or at least

reduce the uncertainties clinicians face when diagnosing patients, for example, during a colonoscopy procedure or when assessing a patient's dementia state. While uncertainties in interacting with AI systems are not new [11], both our cases displayed uncertainties as a result of whether the included dataset was, in fact, representative or complete. Such concerns can undermine clinicians' trust in the system or result in incorrect diagnoses.

At the same time, the cases were distinctly different in the types of training data used. In the colonoscopy case, training data was homogeneous and was collected (recorded from patient footage), labelled, and entered into the system at one institution. The dementia data was much more heterogeneous, with different data elements being brought together through international collaboration.

Researchers should consider diverse diagnostic approaches using different measures for AI-enabled decision support systems. When designing and employing AI technology in decision support situations, researchers and practitioners should be aware of different diagnosis methods and how these differences affect transforming real-world information into digital data.

5.1.4 Overcoming Algorithm Challenges.

Ensuring representative training data. Documenting the source of training data is an essential step toward identifying undesired biases. Accurate documentation furthermore provides ways to interrogate the characteristics of the dataset(s). One example of achieving this is 'The Label' proposed by Holland et al. [45], a diagnostic framework to explore the various ingredients of a training dataset before its use in AI model development. Another example is the 'Datasheets for Datasets' project by Gebru et al. [36], which proposes that each dataset is accompanied by documentation, including *inter alia* its motivation, data collection process, and recommended use. Addressing the same concern, Mitchell et al. developed the Model Cards, short reports about a trained model's characteristics, disclosing intended use and performance evaluation [74]. In addition to ensuring accurate and complete documentation of the dataset used, assessing the validity of the dataset to the target context and population is essential for reliability and end-user trust.

Providing inspectable and adaptable algorithms. AI-enabled decision support systems should be sufficiently flexible to allow end-users to steer recommendations to align with their current task. This can be achieved by, for example, enabling the user to specify the elements most relevant to their current task and updating the AI's assessment accordingly (see *e.g.* [16]). Allowing users to update recommendations by adding or removing specific input variables (*e.g.*, family health history) can enhance an algorithm's inspectability.

Ensuring computation is relevant to the task at hand. More broadly, the lack of a user-centred approach can result in digital systems that fail to meet end-user needs [33]. These systems often fail to be adopted, as they do not provide sufficient value to the intended target group or are too cumbersome to be used in practice. A thorough understanding of (clinical) end-user needs prior to algorithm development will help to overcome these challenges and ensure that the AI-based support offered is relevant to the user's task.

Ensuring replicability and generalisability. While it is impossible to perfectly predict the behaviour of AI-enabled decision support systems across future cases, which are by definition partially unknown, it is paramount to assess the system's replicability. One approach is to verify the AI's assessment with (clinical) experts, using prior clinical assessments of patients as a ground truth against which to compare. A well-known and highly recommended approach is the use of 'training', 'validation', and 'test' data sets. While developing the AI system, it is vital that not all of the data is used for training, as it raises the risk of overfitting on a specific dataset without being able to validate its performance on 'new' observations. As such, the widespread recommendation is to use a training set for training, a validation set to

tune the AI's parameters, and a test set to evaluate an algorithm on data that it has not been trained on to reduce the potential for biases [89].

5.1.5 Presentation. The last stage of MAP considers the presentation of algorithmic outcomes and enables the clinician to infer the real-world meaning of the health situation or patient. We found that our cases shared similarities in augmentation, while they differed in diagnosis characteristics. Both cases involved decision support for clinicians by augmenting existing work practices.

At the same time, the cases differed in how the AI system was used in the real world, relating to the temporal dimension of diagnosis within colonoscopy and dementia. There is little time for debate and reflection within colonoscopy, as the diagnosis is carried out directly and repeatedly throughout the procedure. The dementia case, on the other hand, provided an example of a medical case in which there is more room for reflection and involvement of other clinical specialities. Hence, the diagnosis usually takes place over an extended period.

5.1.6 Overcoming Presentation Challenges.

Ensuring the interpretability of the results. A key criterion for any successful AI-enabled decision support system is that it allows the end-user to put the algorithm's outcome to good use. Here, the presentation does not need to be an exact replication of the clinical reality or the AI's inner calculations. Instead, designers should provide cues (*e.g.*, textual or visual descriptors) that enable the end-user to convert the results into actionable next steps. Here, cues often provide a summarisation or subset of the algorithm's results. Furthermore, interpretability often goes beyond individual roles and specific moments in time, but instead should support repeated reflection and group reasoning across disciplinary boundaries. This embraces the collaborative and iterative nature of clinical decision-making [72].

Adjusting to task context. Integrating with existing real-world practices is vital, stressing the need for a human-centred understanding of the context in which the system is deployed. We found that close involvement of the intended end-users, such as through co-design or participatory design processes, is key for understanding the task context. While this sometimes leads to subtle changes in the presentation of the AI recommendations, at other times we found that design decisions that go against established design norms are required to integrate the AI system into the context satisfactorily. For example, to overcome the fact that AI systems can miss vital information, it may be necessary to display all task-related information to the user rather than displaying only what is marked as relevant by the AI system.

Supporting collaboration between team members. Existing research on Human-AI collaboration largely ignores the fact that professionals often operate in a team, instead framing end-users as soloists with full autonomy and direction over a given task. In collaborative work environments, however, tasks and task responsibilities can change frequently and unexpectedly between individuals. AI support systems should, therefore, support multi-user interaction with the AI system, allowing multiple users to (simultaneously) provide input and request support during cooperative tasks (*e.g.*, surgery). Further, to support the coordination of tasks between team members [68], AI-enabled decision support systems should maintain an updated presentation of the relevant interdependencies between sub-tasks. This can, for example, include a live system status, last decisions made by either human or AI system, or the currently relevant AI recommendations.

Ensuring timeliness of presentation. In designing and studying AI-enabled CDSS, there is a clear need to align the presentation of AI recommendations according to the clinical pertinence. This requires careful consideration of the timeliness of the information presented and an understanding of the information desired at the moment by the clinical staff. We therefore urge designers to thoroughly map the different possible interactions. Here, established techniques

from service design – in which multiple actors engage with information across numerous points in time – could provide a practical guide. For example, 'service blueprints' is a widely used technique to map and visualise user journeys [95].

Minimising confirmation bias and automation complacency. Human biases, such as confirmation bias, can be remedied by implementing bias mitigation strategies in designing AI support systems. Strategies such as 'multiple explanations', in which people consider plausible alternative outcomes for an event, have been shown to debias participant judgements [44]. While the design and integration of these bias mitigation strategies for AI-enabled systems is an open research question, with the suitability of specific mitigation strategies dependent on task context, they provide a possible way forward for designers and researchers concerned with confirmation bias. In addressing concerns related to automation complacency, prior work highlights the value of enhanced quality and safety control measures for different clinical domains [123].

5.2 Building and Maintaining Appropriate User Trust

The successful adoption of AI technology in the clinical domain is dependent on end-user trust. A recent meta-analysis on trust in AI found that trust can be affected by a wide range of factors, related to the user (*e.g.*, personality traits, expertise), the AI (*e.g.*, performance, behaviour), or the context (*e.g.*, risk, communication) [56]. In the colonoscopy case, this wide range of trust-affecting factors was evident. Different roles (*e.g.*, consultant, nurse endoscopist) highlighted various factors which affect their trust towards the AI, and AI failures were assessed differently in different clinical contexts. Alexandra et al. found that trust is not only affected by AI performance but also by factors such as appearance and usability [56]. Barriers to adopting AI systems are therefore highly diverse, including ethical, legal, user experience, and medical aspects. By carefully considering the design and integration of AI systems into clinical workflows, HCI can help to overcome some of these barriers.

Researchers and practitioners have long highlighted the limited explainability, interpretability, and ambiguous accountability of decision-making as significant concerns relating to the use of AI in medicine [3]. Such concerns have been raised from the perspectives of professional healthcare staff [99, 107], patients [106, 107] and users of assistive technology [93]. These rising concerns have led to the creation of numerous principles and guidelines for designing AI systems. However, such high-level principles alone cannot guarantee the development of AI that aligns with medicine's professional, ethical, and regulatory standards. Mittelstadt proposes various pathways towards the development of ethical AI in medicine, including the "*pursuing of ethics as a process, not technological solutionism*" [75]. Such a viewpoint requires a thorough understanding of the actual deployment and use of technology in the work environment, an area in which HCI and related disciplines have built extensive experience [33]. This holds especially true in environments in which agents are expected to cooperate with multiple individuals, a currently underexplored [92] but relevant concern for Human-AI interaction in clinical contexts. For example, in dementia care, clinicians place a high level of trust in colleagues from different disciplines during medical assessment and intervention. Consequently, even if AI systems can explain why a specific recommendation is made, these explanations may not be satisfactory to all members of an interdisciplinary team if they assume high levels of particular domain knowledge. Aligning with Mittelstadt's argument of moving beyond high-level principles [75], we argue that technical fixes alone cannot remove all AI-related concerns. Only by aligning AI systems with the challenges and needs of users can we develop AI systems that function in the real world.

By applying the MAP model during the design of AI-enabled systems, researchers, designers, and developers can contribute to the goal of trust-building in three distinct ways.

First, by making explicit to both system designers and end-users which parameters are used for interpreting a real-world phenomenon, we ensure transparency in how the AI's decision-making comes about. This contributes to the user's understanding of the inner workings of the system (*i.e.*, mental model as intended by Smith & Koppel [96]: "*the way clinicians internally represent and then reason about actions in their clinical world*"), which in turn can enable users to build an understanding of expected and actual AI behaviour during a given task. A lack of transparency in decision-making has been shown to negatively impact user acceptance [20] and satisfaction [59]. To address this, Eiband et al. argue for a participatory process to guide the development of transparent interfaces [29]. Here, the MAP model can help to outline the various parameters and processes leading to the decision outcome. Similarly, the MAP model can assist researchers by informing the use of existing methods for studying trust, such as elicitation cards [93] and focus groups [107]. As indicated by prior work [56], the factors affecting trust are diverse and context-dependent. The MAP model can therefore assist in delimiting the relevant factors to the case being studied.

Second, building on existing medical practices when constructing algorithms can alleviate concerns surrounding the use of automated classification algorithms. Here, the inspectability of the algorithm serves to confirm alignment with proven medical practice and the representativeness of the training dataset.

Third, end-users will only accept new technology if it integrates well with existing work practices. For instance, being presented with an overabundance of false positives results not only in annoyance due to the additional work but ultimately leads to distrust in the system's capabilities [103]. Similarly, adding distractions or delays to the user achieving their goal due to poor contextual fit will result in the abandonment of the AI-enabled decision support system.

5.3 Limitations & Future Work

We recognise several limitations in our work. First, the MAP model presented here covers many everyday situations in which AI systems are deployed in healthcare, and supports reasoning about users' interactions with them. Our case studies demonstrate distinct usage scenarios of AI, one during medical procedures and one in clinical case discussions and investigations. Still, particular kinds of innovative systems may not so readily fit the presented model. While we have not identified such cases, this does not mean they do not exist. Further, we highlight that we have focused exclusively on the usage of AI systems in clinical practice. The role of AI-enabled systems in the education or training of clinical staff is an essential area for future work.

Second, our model focuses on human-in-the-loop systems, particularly within CDSS, rather than AI systems that operate autonomously. Autonomous systems in the medical domain, sometimes called closed-loop systems, have been contentious due to the potential negative impact of autonomous AI systems on patient safety. Regulatory organisations, such as the FDA (U.S. Food and Drug Administration), have traditionally not approved systems that exclude the human operator [22]. While increasingly intelligent AI systems might one day challenge this paradigm, we believe that human-in-the-loop systems will continue to play a dominant role in medical decision-making due to the high stakes, need for accountability, and ability to communicate and weigh decisions with patients and clinical team members. Supporting end-user interaction (whether patient, clinician, or other stakeholders) with fully autonomous systems is likely to require different interaction paradigms, as can be seen, for example, in the discussion on autonomous driving.

Finally, in this article, we consider the interaction with AI systems from a clinician's point of view. With the increasing use of AI in medical practice, patients will also be confronted with AI systems more frequently as part of their medical interactions. As such, HCI researchers should carefully consider supporting patients and their caregivers in interacting with AI-enabled systems. These requirements are likely to differ due to their typically limited domain knowledge compared to medical professionals, a desire and need to balance medical decisions with other aspects of life, widely

varying technical skills and proficiency, and other factors. We, therefore, consider the design of solutions that support patients in interacting with AI systems as a critical area for future work in HCI to support and maintain the vital practice of shared patient-clinician decision-making [30].

6 CONCLUSION

The challenges of medical practice, ranging from 'first, do no harm' ('*primum non nocere*' in the Hippocratic Oath) to patients' right to privacy and access to appropriate care, demands careful consideration of the introduction of AI-enabled systems in healthcare. Based on the rich literature that has emerged in this space, combined with our prior experiences designing and evaluating AI-enabled systems for endoscopy and dementia, we have identified a three-stage procedure through which AI systems interpret medical reality. We formalise this procedure in the MAP model, in which we distinguish three stages through which medical observations are interpreted and handled by AI systems. For each stage, *Measurement, Algorithm, Presentation*, we have inductively identified challenges that apply to the design of Human-Centred AI systems, focusing on CDSS. If such systems are to be accepted and effective in clinical practice, these challenges need to be addressed. While the list of challenges may not yet be complete, it provides an evidence-based checklist to guide researchers and practitioners in the design and evaluation of AI-enabled systems in healthcare.

ACKNOWLEDGMENTS

The authors would like to thank all the clinicians who took part in the studies for their valuable contributions. This work is supported by the EXPLAIN-ME project of Digital Research Centre Denmark (DIREC) under Innovation Fund Denmark, the Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS) (NS/A000050/1), the EPSRC CDT in Medical Imaging (EP/L016478/1), and the industrial partner icometrix. This work also received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 666992.

REFERENCES

[1] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 3. https://doi.org/10.1145/3290605.3300233

[2] John R Anderson and Christian J Lebiere. 2014. *The atomic components of thought.* Psychology Press.

[3] Boris Babic, Sara Gerke, Theodoros Evgeniou, and I. Glenn Cohen. 2021. Beware explanations from AI in health care. *Science* 373, 6552 (2021), 284–286. https://doi.org/10.1126/science.abg1834

[4] Donald P Ballou and Harold L Pazer. 1985. Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems. *Management Science* 31, 2 (1985), 150–162.

[5] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter Lasecki, Dan Weld, and Eric Horvitz. 2019. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. In *HCOMP*. AAAI.

[6] David W. Bates, David Levine, Ania Syrowatka, Masha Kuznetsova, Kelly Jean Thomas Craig, Angela Rui, Gretchen Purcell Jackson, and Kyu Rhee. 2021. The potential of artificial intelligence to improve patient safety: a scoping review. *npj Digital Medicine* 4, 1 (2021), 54. https://doi.org/10.1038/s41746-021-00423-6

[7] Maura Bellio. 2021. *Translating Predictive Models for Alzheimer's Disease to Clinical Practice: User Research, Adoption Opportunities, and Conceptual Design of a Decision Support Tool.* Ph. D. Dissertation. UCL (University College London).

[8] Maura Bellio, Dominic Furniss, Neil P Oxtoby, Sara Garbarino, Nicholas C Firth, Annemie Ribbens, Daniel C Alexander, and Ann Blandford. 2021. Opportunities and Barriers for Adoption of a Decision-Support Tool for Alzheimer's Disease. *ACM Transactions on Computing for Healthcare* 2, 4 (2021), 1–19. https://doi.org/10.1145/3462764

[9] Maura Bellio, Neil P Oxtoby, Annemie Ribbens, Daniel C Alexander, and Ann Blandford. 2020. Testing the conceptual design of icompass: a new clinical decision-support tool to guide clinicians through Alzheimer's Disease markers' evolution. In *2020 Alzheimer's Association International Conference*. ALZ.

[10] Maura Bellio, Neil P Oxtoby, Zuzana Walker, Susie Henley, Annemie Ribbens, Ann Blandford, Daniel C Alexander, and Keir XX Yong. 2020. Analyzing large Alzheimer's disease cognitive datasets: Considerations and challenges. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease*

Monitoring 12, 1 (2020), e12135. https://doi.org/10.1002/dad2.12135

[11] Jesse Josua Benjamin, Arne Berger, Nick Merrill, and James Pierce. 2021. Machine Learning Uncertainty as a Design Material: A Post-Phenomenological Inquiry. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, Article 171, 14 pages. https://doi.org/10.1145/3411764.3445481

[12] Ann Blandford. 2001. Intelligent interaction design: the role of human-computer interaction research in the design of intelligent systems. *Expert Systems* 18, 1 (2001), 3–18. https://doi.org/10.1111/1468-0394.00151

[13] Ann Blandford, Richard Butterworth, and Paul Curzon. 2004. Models of interactive systems: a case study on programmable user modelling. *International Journal of Human-Computer Studies* 60, 2 (2004), 149–200. https://doi.org/10.1016/j.ijhcs.2003.08.004

[14] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 77–91.

[15] Michael D Byrne and Alex Kirlik. 2005. Using computational cognitive modeling to diagnose possible sources of aviation error. *The international journal of aviation psychology* 15, 2 (2005), 135–155.

[16] Carrie J. Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S. Corrado, Martin C. Stumpe, and Michael Terry. 2019. Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 1–14. https://doi.org/10.1145/3290605.3300234

[17] José Creissac Campos, Gavin Doherty, and Michael D. Harrison. 2014. Analysing interactive devices based on information resource constraints. *International Journal of Human-Computer Studies* 72, 3 (2014), 284–297. https://doi.org/10.1016/j.ijhcs.2013.10.005

[18] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-Day Readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, 1721–1730. https://doi.org/10.1145/2783258.2788613

[19] Robert Challen, Joshua Denny, Martin Pitt, Luke Gompels, Tom Edwards, and Krasimira Tsaneva-Atanasova. 2019. Artificial intelligence, bias and clinical safety. *BMJ Quality & Safety* 28, 3 (2019), 231–237. https://doi.org/10.1136/bmjqs-2018-008370

[20] Henriette Cramer, Vanessa Evers, Satyan Ramlal, Maarten van Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob Wielinga. 2008. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction* 18, 5 (2008), 455. https://doi.org/10.1007/s11257-008-9051-3

[21] Pat Croskerry. 2009. A Universal Model of Diagnostic Reasoning. *Academic Medicine* 84, 8 (2009). https://doi.org/10.1097/ACM.0b013e3181ace703

[22] M.L. Cummings and David Britton. 2020. Regulating safety-critical autonomous systems: past, present, and future perspectives. In *Living with Robots*, Richard Pak, Ewart J. de Visser, and Ericka Rovira (Eds.). Academic Press, 119–140. https://doi.org/10.1016/B978-0-12-815367-3.00006-2

[23] Adrienne Curry, Peter Flett, and Ivan Hollingsworth. 2006. *Managing information & systems: The business perspective*. Routledge. 1–5 pages.

[24] Sabyasachi Dash, Sushil Kumar Shakyawar, Mohit Sharma, and Sandeep Kaushik. 2019. Big data in healthcare: management, analysis and future prospects. *Journal of Big Data* 6, 1 (2019), 1–25. https://doi.org/10.1186/s40537-019-0217-0

[25] Jeffrey De Fauw, Joseph R. Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, George van den Driessche, Balaji Lakshminarayanan, Clemens Meyer, Faith Mackinder, Simon Bouton, Kareem Ayoub, Reena Chopra, Dominic King, Alan Karthikesalingam, Cían O. Hughes, Rosalind Raine, Julian Hughes, Dawn A. Sim, Catherine Egan, Adnan Tufail, Hugh Montgomery, Demis Hassabis, Geraint Rees, Trevor Back, Peng T. Khaw, Mustafa Suleyman, Julien Cornebise, Pearse A. Keane, and Olaf Ronneberger. 2018. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine* 24, 9 (2018), 1342–1350. https://doi.org/10.1038/s41591-018-0107-6

[26] S. E. Dilsizian and E. L. Siegel. 2014. Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. *Curr Cardiol Rep* 16, 1 (2014), 441. https://doi.org/10.1007/s11886-013-0441-8

[27] Alan Dix. 2007. Designing for appropriation. In *Proceedings of HCI 2007 The 21st British HCI Group Annual Conference University of Lancaster, UK 21*. 1–4.

[28] Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. 2017. UX Design Innovation: Challenges for Working with Machine Learning as a Design Material. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 278–288. https://doi.org/10.1145/3025453.3025739

[29] Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. 2018. Bringing Transparency Design into Practice. In *23rd International Conference on Intelligent User Interfaces (IUI '18)*. Association for Computing Machinery, 211–223. https://doi.org/10.1145/3172944.3172961

[30] Glyn Elwyn, Dominick Frosch, Richard Thomson, Natalie Joseph-Williams, Amy Lloyd, Paul Kinnersley, Emma Cording, Dave Tomson, Carole Dodd, Stephen Rollnick, Adrian Edwards, and Michael Barry. 2012. Shared decision making: a model for clinical practice. *Journal of General Internal Medicine* 27, 10 (2012), 1361–1367. https://doi.org/10.1007/s11606-012-2077-6

[31] European Commission. 2021. PDiagnostic accuracy of digital screening mame European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. COM/2021/206. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206.

[32] Samuel G. Finlayson, John D. Bowers, Joichi Ito, Jonathan L. Zittrain, Andrew L. Beam, and Isaac S. Kohane. 2019. Adversarial attacks on medical machine learning. *Science* 363, 6433 (2019), 1287–1289. https://doi.org/10.1126/science.aaw4399

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

[34] Susanne Gaube, Harini Suresh, Martina Raue, Alexander Merritt, Seth J. Berkowitz, Eva Lermer, Joseph F. Coughlin, John V. Guttag, Errol Colak, and Marzyeh Ghassemi. 2021. Do as AI say: susceptibility in deployment of clinical decision-aids. *npj Digital Medicine* 4, 1 (2021), 31. https://doi.org/10.1038/s41746-021-00385-9

[35] Andrew J. Gawron, Annapoorani Veerappan, and Rajesh N. Keswani. 2014. High success rate of repeat colonoscopy with standard endoscopes in patients referred for prior incomplete colonoscopy. *BMC Gastroenterology* 14, 1 (2014), 56. https://doi.org/10.1186/1471-230X-14-56

[36] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for Datasets. *Commun. ACM* 64, 12 (2021), 86–92. https://doi.org/10.1145/3458723

[37] Benjamin A Goldstein, Ann Marie Navar, Michael J Pencina, and John Ioannidis. 2017. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association* 24, 1 (2017), 198–208. https://doi.org/10.1093/jamia/ocw042

[38] Dennis S Gouran. 1973. Group communication: Perspectives and priorities for future research. *Quarterly Journal of Speech* 59, 1 (1973), 22–29.

[39] Jonathan Grudin. 2006. Turing Maturing: The Separation of Artificial Intelligence and Human-Computer Interaction. *Interactions* 13, 5 (2006), 54–57. https://doi.org/10.1145/1151314.1151346

[40] Jonathan Grudin. 2009. AI and HCI: Two Fields Divided by a Common Focus. *AI Magazine* 30, 4 (2009), 48. https://doi.org/10.1609/aimag.v30i4.2271

[41] Magali Haas, Diane Stephenson, Klaus Romero, Mark Forrest Gordon, Neta Zach, and Hugo Geerts. 2016. Big data to smart data in Alzheimer's disease: Real-world examples of advanced modeling and simulation. *Alzheimer's & Dementia* 12, 9 (2016), 1022–1030. https://doi.org/10.1016/j.jalz.2016.05.005

[42] Daniel A. Hashimoto, Guy Rosman, Daniela Rus, and Ozanan R. Meireles. 2018. Artificial Intelligence in Surgery: Promises and Perils. *Annals of Surgery* 268, 1 (2018), 70–76. https://doi.org/10.1097/SLA.0000000000002693

[43] Richard Heeks. 2006. Health information systems: Failure, success and improvisation. *International Journal of Medical Informatics* 75, 2 (2006), 125–137. https://doi.org/10.1016/j.ijmedinf.2005.07.024

[44] Edward R. Hirt and Keith D. Markman. 1995. Multiple explanation: A consider-an-alternative strategy for debiasing judgments. *Journal of Personality and Social Psychology* 69, 6 (1995), 1069–1086. https://doi.org/10.1037/0022-3514.69.6.1069

[45] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The dataset nutrition label: A framework to drive higher data quality standards. *arXiv preprint arXiv:1805.03677* (2018).

[46] Andreas Holzinger, Chris Biemann, Constantinos S. Pattichis, and Douglas B. Kell. 2017. What do we need to build explainable AI systems for the medical domain? arXiv:1712.09923 [cs.AI]

[47] Jan Horsky, Gordon D. Schiff, Douglas Johnston, Lauren Mercincavage, Douglas Bell, and Blackford Middleton. 2012. Interface design principles for usable decision support: A targeted review of best practices for clinical prescribing interventions. *Journal of Biomedical Informatics* 45, 6 (2012), 1202–1216. https://doi.org/10.1016/j.jbi.2012.09.002

[48] Ahmed Hosny, Chintan Parmar, John Quackenbush, Lawrence H. Schwartz, and Hugo J. W. L. Aerts. 2018. Artificial intelligence in radiology. *Nature Reviews Cancer* 18, 8 (2018), 500–510. https://doi.org/10.1038/s41568-018-0016-5

[49] Yipeng Hu, Joseph Jacob, Geoffrey JM Parker, David J Hawkes, John R Hurst, and Danail Stoyanov. 2020. The challenges of deploying artificial intelligence models in a rapidly evolving pandemic. *Nature Machine Intelligence* 2, 6 (2020), 298–300. https://doi.org/10.1038/s42256-020-0185-2

[50] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, 560–575. https://doi.org/10.1145/3442188.3445918

[51] Sophie F. Jentzsch, Sviatlana Höhn, and Nico Hochgeschwender. 2019. Conversational Interfaces for Explainable AI: A Human-Centred Approach. In *Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, Davide Calvaresi, Amro Najjar, Michael Schumacher, and Kary Främling (Eds.). Springer International Publishing, 77–92. https://doi.org/10.1007/978-3-030-30391-4_5

[52] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. 2020. Kvasir-seg: A segmented polyp dataset. In *International Conference on Multimedia Modeling*. Springer, 451–462.

[53] Fei Jiang, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, Yilong Wang, Qiang Dong, Haipeng Shen, and Yongjun Wang. 2017. Artificial intelligence in healthcare: past, present and future. *Stroke and Vascular Neurology* 2, 4 (2017), 230–243. https://doi.org/10.1136/svn-2017-000101

[54] Kenneth Jung, Sehj Kashyap, Anand Avati, Stephanie Harman, Heather Shaw, Ron Li, Margaret Smith, Kenny Shum, Jacob Javitz, Yohan Vetteth, Tina Seto, Steven C Bagley, and Nigam H Shah. 2020. A framework for making predictive models useful in practice. *Journal of the American Medical Informatics Association* 28, 6 (2020), 1149–1158. https://doi.org/10.1093/jamia/ocaa318

[55] P. Kandel and M. B. Wallace. 2019. Should We Resect and Discard Low Risk Diminutive Colon Polyps. *Clinical Endoscopy* 52, 3 (2019), 239–246. https://doi.org/10.5946/ce.2018.136

[56] Alexandra D. Kaplan, Theresa T. Kessler, J. Christopher Brill, and P. A. Hancock. 2021. Trust in Artificial Intelligence: Meta-Analytic Findings. *Human Factors* (2021), 00187208211013988. https://doi.org/10.1177/00187208211013988

[57] Anastasiya Kiseleva. 2020. AI as a Medical Device: Is It Enough to Ensure Performance Transparency and Accountability in Healthcare? *European Pharmaceutical Law Review* 1 (2020).

[58] Gary A Klein. 2017. *Sources of power: How people make decisions*. MIT press.

[59] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell Me More? The Effects of Mental Model Soundness on Personalizing an Intelligent Agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. Association for Computing Machinery, 1–10. https://doi.org/10.1145/2207676.2207678

[60] David R. Large, Gary Burnett, and Leigh Clark. 2019. Lessons from Oz: Design Guidelines for Automotive Conversational User Interfaces. In *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications: Adjunct Proceedings (AutomotiveUI '19)*. Association for Computing Machinery, 335–340. https://doi.org/10.1145/3349263.3351314

[61] Kate Laver, Monica Cations, Gorjana Radisic, Lenore De La Perrelle, Richard Woodman, Janna Anneke Fitzgerald, Susan Kurrle, Ian D Cameron, Craig Whitehead, Jane Thompson, et al. 2020. Improving adherence to guideline recommendations in dementia care through establishing a quality improvement collaborative of agents of change: an interrupted time series study. *Implementation Science Communications* 1, 1 (2020), 1–11. https://doi.org/10.1186/s43058-020-00073-x

[62] Chang Kyun Lee, Dong Il Park, Suck-Ho Lee, Young Hwangbo, Chang Soo Eun, Dong Soo Han, Jae Myung Cha, Bo-In Lee, and Jeong Eun Shin. 2011. Participation by experienced endoscopy nurses increases the detection rate of colon polyps during a screening colonoscopy: a multicenter, prospective, randomized study. *Gastrointestinal Endoscopy* 74, 5 (2011), 1094–1102. https://doi.org/10.1016/j.gie.2011.06.033

[63] Constance D Lehman, Robert D Wellman, Diana SM Buist, Karla Kerlikowske, Anna NA Tosteson, Diana L Miglioretti, and Breast Cancer Surveillance Consortium. 2015. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Internal Medicine* 175, 11 (2015), 1828–1837. https://doi.org/10.1001/jamainternmed.2015.5231

[64] Jill Fain Lehman, John Laird, and Paul Rosenbloom. 2006. A gentle introduction to soar, an architecture for human cognition: 2006 update. *University of Michigan* (2006), 1–37.

[65] Ron C. Li, Steven M. Asch, and Nigam H. Shah. 2020. Developing a delivery science for artificial intelligence in healthcare. *npj Digital Medicine* 3, 1 (2020), 107. https://doi.org/10.1038/s41746-020-00318-y

[66] Q. Vera Liao, Daniel M. Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. *CoRR* abs/2001.02478 (2020). http://arxiv.org/abs/2001.02478

[67] Lucy C Maling, Christian EB Gray-Stephens, Khalid Malik-Tabassum, Oliver JF Weiner, Matthew R Marples, Giles P Faria, and Rory G Middleton. 2021. The National Hip Fracture Database is only as good as the data we feed it - significant inaccuracy demonstrated and how to improve it. *Injury* 52, 4 (2021), 894–897. https://doi.org/10.1016/j.injury.2020.10.079

[68] Thomas W. Malone and Kevin Crowston. 1990. What is Coordination Theory and How Can It Help Design Cooperative Work Systems?. In *Proceedings of the 1990 ACM Conference on Computer-Supported Cooperative Work (CSCW '90)*. Association for Computing Machinery, 357–370. https://doi.org/10.1145/99332.99367

[69] David Manning, Susan Ethell, Tim Donovan, and Trevor Crawford. 2006. How do radiologists do it? The influence of experience and training on searching for chest nodules. *Radiography* 12, 2 (2006), 134–142. https://doi.org/10.1016/j.radi.2005.02.003

[70] Markets and Markets. 2020. Artificial Intelligence in Healthcare Market with Covid-19 Impact Analysis by Offering, Technology, End-Use Application, End User and Region - Global Forecast to 2026.

[71] Harriet Martineau. 1859. *England and her Soldiers*.

[72] Anne Miller, Jejo D. Koola, Michael E. Matheny, Julie H. Ducom, Jason M. Slagle, Erik J. Groessl, Freneka F. Minter, Jennifer H. Garvin, Matthew B. Weinger, and Samuel B. Ho. 2018. Application of contextual design methods to inform targeted clinical decision support interventions in sub-specialty care environments. *International Journal of Medical Informatics* 117 (2018), 55–65. https://doi.org/10.1016/j.ijmedinf.2018.05.005

[73] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[74] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, 220–229. https://doi.org/10.1145/3287560.3287596

[75] Brent Mittelstadt. 2019. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence* 1, 11 (2019), 501–507. https://doi.org/10.1038/s42256-019-0114-4

[76] Y. Mori, S. E. Kudo, T. M. Berzin, M. Misawa, and K. Takeda. 2017. Computer-aided diagnosis for colonoscopy. *Endoscopy* 49, 8 (2017), 813–819. https://doi.org/10.1055/s-0043-109430

[77] Miho Nakajima, L.Ian Schmitt, and Michael M. Halassa. 2019. Prefrontal Cortex Regulates Sensory Filtering through a Basal Ganglia-to-Thalamus Pathway. *Neuron* 103, 3 (2019), 445–458.e10. https://doi.org/10.1016/j.neuron.2019.05.026

[78] Nariman Noorbakhsh-Sabet, Ramin Zand, Yanfei Zhang, and Vida Abedi. 2019. Artificial Intelligence Transforms the Future of Health Care. *The American Journal of Medicine* 132, 7 (2019), 795–801. https://doi.org/10.1016/j.amjmed.2019.01.017

[79] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. 2019. InterpretML: A Unified Framework for Machine Learning Interpretability. arXiv:1909.09223 [cs.LG]

[80] Donald A Norman. 1983. Some observations on mental models. *Mental models* 7, 112 (1983), 7–14.

[81] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453. https://doi.org/10.1126/science.aax2342

[82] Neil P Oxtoby and Daniel C Alexander. 2017. Imaging plus X: multimodal models of neurodegenerative disease. *Current Opinion in Neurology* 30, 4 (2017), 371. https://doi.org/10.1097/WCO.0000000000000460

[83] Charles Pavitt. 2014. An interactive input–process–output model of social influence in decision-making groups. *Small Group Research* 45, 6 (2014), 704–730.

[84] Ronald Carl Petersen, PS Aisen, Laurel A Beckett, MC Donohue, AC Gamst, Danielle J Harvey, CR Jack, WJ Jagust, LM Shaw, AW Toga, et al. 2010. Alzheimer's disease neuroimaging initiative (ADNI): clinical characterization. *Neurology* 74, 3 (2010), 201–209. https://doi.org/10.1212/WNL.0b013e3181cb3e25

[85] Joshua Pink, John O'Brien, Louise Robinson, and Damien Longson. 2018. Dementia: assessment, management and support: summary of updated NICE guidance. *BMJ* 361 (2018). https://doi.org/10.1136/bmj.k2438

[86] C. Pox, W. Schmiegel, and M. Classen. 2007. Current status of screening colonoscopy in Europe and in the United States. *Endoscopy* 39, 2 (2007), 168–173. https://doi.org/10.1055/s-2007-966182

[87] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. 2018. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine* 1, 1 (2018), 1–10. https://doi.org/10.1038/s41746-018-0029-1

[88] Scott Reeves, Simon Lewin, Sherry Espin, and Merrick Zwarenstein. 2011. *Interprofessional Teamwork for Health and Social Care*. John Wiley & Sons. https://doi.org/10.1002/9781444325027

[89] Brian D Ripley. 2007. *Pattern recognition and neural networks*. Cambridge University Press.

[90] Peter Safar, Torrey C. Brown, Warren J. Holtey, and Robert J. Wilder. 1961. Ventilation and Circulation with Closed-Chest Cardiac Massage in Man. *JAMA* 176, 7 (1961), 574–576. https://doi.org/10.1001/jama.1961.03040200010003

[91] Kritika Samsi and Jill Manthorpe. 2014. Care pathways for dementia: current perspectives. *Clinical Interventions in Aging* 9 (2014), 2055. https://doi.org/10.2147/CIA.S70628

[92] Isabel Schwaninger, Geraldine Fitzpatrick, and Astrid Weiss. 2019. Exploring Trust in Human-Agent Collaboration. In *Proceedings of 17th European Conference on Computer-Supported Cooperative Work*. European Society for Socially Embedded Technologies (EUSSET). https://doi.org/10.18420/ecscw2019_ep08

[93] Isabel Schwaninger, Florian Güldenpfennig, Astrid Weiss, and Geraldine Fitzpatrick. 2021. What Do You Mean by Trust? Establishing Shared Meaning in Interdisciplinary Design for Assistive Technology. *International Journal of Social Robotics* 13, 8 (01 Dec 2021), 1879–1897. https://doi.org/10.1007/s12369-020-00742-w

[94] Ben Shneiderman. 2020. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction* 36, 6 (2020), 495–504. https://doi.org/10.1080/10447318.2020.1741118

[95] Lynn Shostack. 1984. Designing services that deliver. *Harvard Business Review* 62, 1 (1984), 133–139.

[96] Sean W Smith and Ross Koppel. 2014. Healthcare information technology's relativity problems: a typology of how patients' physical reality, clinicians' mental models, and healthcare information technology differ. *Journal of the American Medical Informatics Association* 21, 1 (2014), 117–131. https://doi.org/10.1136/amiajnl-2012-001419

[97] Anja Thieme, Danielle Belgrave, and Gavin Doherty. 2020. Machine Learning in Mental Health: A Systematic Review of the HCI Literature to Support the Development of Effective and Implementable ML Systems. *ACM Trans. Comput.-Hum. Interact.* 27, 5, Article 34 (2020), 53 pages. https://doi.org/10.1145/3398069

[98] UK Space Agency. 2019. Space tech to fight bowel cancer and exposure to air pollution. https://www.gov.uk/government/news/space-tech-to-fight-bowel-cancer-and-exposure-to-air-pollution

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

[100] Niels van Berkel, Matthew J. Clarkson, Guofang Xiao, Eren Dursun, Moustafa Allam, Brian R. Davidson, and Ann Blandford. 2020. Dimensions of ecological validity for usability evaluations in clinical settings. *Journal of Biomedical Informatics* 110 (2020), 103553. https://doi.org/10.1016/j.jbi.2020.103553

[101] Niels van Berkel, Jorge Goncalves, Simo Hosio, Zhanna Sarsenbayeva, Eduardo Velloso, and Vassilis Kostakos. 2020. Overcoming compliance bias in self-report studies: A cross-study analysis. *International Journal of Human-Computer Studies* 134 (2020), 1–12. https://doi.org/10.1016/j.ijhcs.2019.10.003

[102] Niels van Berkel, Jorge Goncalves, Daniel Russo, Simo Hosio, and Mikael B. Skov. 2021. Effect of Information Presentation on Fairness Perceptions of Machine Learning Predictors. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, Article 245, 13 pages. https://doi.org/10.1145/3411764.3445365

[103] Niels van Berkel, Jeremy Opie, Omer F. Ahmad, Laurence Lovat, Danail Stoyanov, and Ann Blandford. 2022. Initial Responses to False Positives in AI-supported Continuous Interactions: A Colonoscopy Case Study. *ACM Transactions on Interactive Intelligent Systems* 12, 1, Article 2 (2022), 18 pages. https://doi.org/10.1145/3480247

[104] Niels van Berkel, Mikael B. Skov, and Jesper Kjeldskov. 2021. Human-AI Interaction: Intermittent, Continuous, and Proactive. *Interactions* 28, 6 (2021), 67–71. https://doi.org/10.1145/3486941

[105] J. C. van Rijn, J. B. Reitsma, J. Stoker, P. M. Bossuyt, S. J. van Deventer, and E. Dekker. 2006. Polyp miss rate determined by tandem colonoscopy: a systematic review. *The American Journal of Gastroenterology* 101, 2 (2006), 343–350. https://doi.org/10.1111/j.1572-0241.2006.00390.x

[106] L. van Velsen, I. Flierman, and M. Tabak. 2021. The formation of patient trust and its transference to online health services: the case of a Dutch online patient portal for rehabilitation care. *BMC Medical Informatics and Decision Making* 21, 1 (2021), 188. https://doi.org/10.1186/s12911-021-01552-4

[107] Lex Van Velsen, Sabine Wildevuur, Ina Flierman, Boris Van Schooten, Monique Tabak, and Hermie Hermens. 2016. Trust in telemedicine portals for rehabilitation care: an exploratory focus group study with patients and healthcare professionals. *BMC Medical Informatics and Decision Making* 16, 1 (27 Jan 2016), 11. https://doi.org/10.1186/s12911-016-0250-2

[108] Sebastian Vollmer, Bilal A Mateen, Gergo Bohner, Franz J Király, Rayid Ghani, Pall Jonsson, Sarah Cumbers, Adrian Jonas, Katherine SL McAllister, Puja Myles, et al. 2020. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *bmj* 368 (2020). https://doi.org/10.1136/bmj.l6927

[109] Brian Wahl, Aline Cossy-Gantner, Stefan Germann, and Nina R Schwalbe. 2018. Artificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings? *BMJ Global Health* 3, 4 (2018). https://doi.org/10.1136/bmjgh-2018-000798

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

[111] Dakuo Wang, Justin D. Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. 2019. Human-AI Collaboration in Data Science: Exploring Data Scientists' Perceptions of Automated AI. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 211 (2019), 24 pages. https://doi.org/10.1145/3359313

[112] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 1–15. https://doi.org/10.1145/3290605.3300831

[113] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. 2020. The What-If Tool: Interactive Probing of Machine Learning Models. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2020), 56–65. https://doi.org/10.1109/TVCG.2019.2934619

[114] Jeannette M. Wing. 2006. Computational Thinking. *Commun. ACM* 49, 3 (2006), 33–35. https://doi.org/10.1145/1118178.1118215

[115] Allison Woodruff, Sarah E. Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A Qualitative Exploration of Perceptions of Algorithmic Fairness. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. 1–14. https://doi.org/10.1145/3173574.3174230

[116] Antje Wulff, Sara Montag, Bianca Steiner, Michael Marschollek, Philipp Beerbaum, André Karch, and Thomas Jack. 2019. CADDIE2—evaluation of a clinical decision-support system for early detection of systemic inflammatory response syndrome in paediatric intensive care: study protocol for a diagnostic study. *BMJ open* 9, 6 (2019), e028953. https://doi.org/10.1136/bmjopen-2019-028953

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

[118] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. Unremarkable AI: Fitting Intelligent Decision Support into Critical, Clinical Decision-Making Processes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 1–11. https://doi.org/10.1145/3290605.3300468

[119] Qian Yang, John Zimmerman, Aaron Steinfeld, Lisa Carey, and James F Antaki. 2016. Investigating the heart pump implant decision process: opportunities for decision support tools to help. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. 4477–4488. https://doi.org/10.1145/2858036.2858373

[120] Ryosuke Yokoi, Yoko Eguchi, Takanori Fujita, and Kazuya Nakayachi. 2021. Artificial Intelligence Is Trusted Less than a Doctor in Medical Treatment Decisions: Influence of Perceived Care and Value Similarity. *International Journal of Human–Computer Interaction* 37, 10 (2021), 981–990. https://doi.org/10.1080/10447318.2020.1861763

[121] Alexandra L Young, Neil P Oxtoby, Pankaj Daga, David M Cash, Nick C Fox, Sebastien Ourselin, Jonathan M Schott, and Daniel C Alexander. 2014. A data-driven model of biomarker changes in sporadic Alzheimer's disease. *Brain* 137, 9 (2014), 2564–2577. https://doi.org/10.1093/brain/awu176

[122] Kun-Hsing Yu, Andrew L. Beam, and Isaac S. Kohane. 2018. Artificial intelligence in healthcare. *Nature Biomedical Engineering* 2, 10 (2018), 719–731. https://doi.org/10.1038/s41551-018-0305-z

[123] Kun-Hsing Yu and Isaac S Kohane. 2019. Framing the challenges of artificial intelligence in medicine. *BMJ Quality & Safety* 28, 3 (2019), 238–241. https://doi.org/10.1136/bmjqs-2018-008551

[124] John Zhang. 2007. Effect of Age and Sex on Heart Rate Variability in Healthy Subjects. *Journal of Manipulative and Physiological Therapeutics* 30, 5 (2007), 374–379. https://doi.org/10.1016/j.jmpt.2007.04.001

[125] Shengbing Zhao, Shuling Wang, Peng Pan, Tian Xia, Xin Chang, Xia Yang, Liliangzi Guo, Qianqian Meng, Fan Yang, Wei Qian, Zhichao Xu, Yuanqiong Wang, Zhijie Wang, Lun Gu, Rundong Wang, Fangzhou Jia, Jun Yao, Zhaoshen Li, and Yu Bai. 2019. Magnitude, Risk Factors, and Factors Associated With Adenoma Miss Rate of Tandem Colonoscopy: A Systematic Review and Meta-analysis. *Gastroenterology* 156, 6 (2019), 1661–1674.e11. https://doi.org/10.1053/j.gastro.2019.01.260