Making AI Work: Designing and Evaluating AI Systems in Healthcare

Niels van Berkel
Department of Computer Science, Aalborg University, Aalborg, Denmark

Learning Objectives

- Understand the field of human-computer interaction (HCI), as well as have an overview of relevant work in this area for further reading.
- Understand some of the factors that support the alignment of AI systems with the needs of medical professionals, patients, and other relevant stakeholders.
- Recognize the importance of stakeholder involvement when designing human-AI systems in healthcare.
- Consider the role and importance of a study's ecological validity, longitudinal assessments, validation across diverse datasets, and iterative development when designing human-AI systems in healthcare.
- Discuss recommendations for the development of AI systems in healthcare, with lessons learned from prior deployments.

41.1 Introduction

Promising stories of AI have set sky-high expectations for its ability to transform medical practice. However, living up to even the more modest expectations will require careful consideration of how medical professionals will interact with this technology and how we can successfully embed AI systems within day-to-day clinical practice. Previous attempts at introducing technology into the medical context, such as electronic health records [33], have not been unanimously successful [23]. This is despite what is often initial optimism regarding a technology's ability to transform healthcare. Similarly, in the recent and ongoing battle against COVID-19, many AI systems were developed to support the diagnosis and triaging of patients, the impact of which was ultimately limited [37]. These challenges highlight the difficulties of designing and evaluating digital systems for use in the healthcare context, which may be further exacerbated if the technology does not meet the needs of medical professionals. Yet, the potential for AI systems to support medical professionals is vast, covering areas such as diagnosis, prognosis, guided surgery, and assisting in the training of medical personnel. This chapter provides a starting point for developers and clinicians interested in the design and evaluation of AI systems in a healthcare context.

The field of human-computer interaction (HCI) has developed methods and guidelines on how to design and evaluate digital systems so that they

AI in Clinical Medicine: A Practical Guide for Healthcare Professionals, First Edition. Edited by Michael F. Byrne, Nasim Parsa, Alexandra T.Greenhill, Daljeet Chahal, Omer Ahmad, and Ulas Bagci.
 Published 2023 by John Wiley & Sons Ltd.
Companion website: www.wiley.com/go/byrne/aiinclinicalmedicine

align with the needs of the system's users. As a research discipline, HCI has been described as sitting at the intersection of computer science and psychology [12]. Combining these two disciplines and following a mindset in which the user of the system is central in developing new technology, HCI has stressed the importance of understanding user needs and the context in which the technology will be used in order to create successful digital systems. More recently, HCI researchers and practitioners have begun to shift their attention and expertise to the design of AI systems. Following the aforementioned focus on the human stakeholders in any technology, this work has been described as human-AI [1] or human-centered AI [43].

In this chapter, we provide concrete recommendations for the development of AI systems in healthcare. HCI and related disciplines have a long history of contributing to the healthcare technology landscape [17]. While AI has only recently matured to be deployed in real-world applications, experiences obtained in the design and evaluation of non-AI based systems provide valuable lessons that similarly apply to AI technology design. Through examples from the literature and based on our own experiences in designing and evaluating AI systems, we will illustrate lessons learned from prior deployments. The following section provides a high-level overview of the contributions and history of HCI-driven work in healthcare. This is followed by this chapter's two primary contributions: recommendations for the *design* and *evaluation* of human-AI systems in healthcare.

How do we design systems that medical professionals want to use in their day-to-day activities? Beyond the critical aspect of ensuring the accuracy of these AI systems, the usability and user experience of this technology are vital in ensuring adoption. Failing to meet these aspects will result in the system being abandoned, regardless of its potential in improving healthcare quality or efficiency. The evaluation of human-AI systems is complex, as system behaviour depends heavily on the input received and subsequent assessment of the algorithm. This makes evaluation less predictable. This level of unpredictability, combined with a medical context in which the evaluation of prototype systems on actual patients is often not considered, raising novel challenges for assessing AI systems. However, evaluation is critical, as it can reveal essential shortcomings in a system prior to eventual deployment.

Finally, we outline open research questions that have not been extensively discussed within the literature but require further investigation. These problems deal with embedding AI technology in a team rather than supporting one individual, involving patients in algorithm-driven decision-making, and effectively sharing tasks between the human operator and the AI system.

41.2 Background in Human-Computer Interaction

Prior to the recent increase in AI-driven systems, HCI researchers and practitioners have already gained a wide range of experience in designing interactive systems. Of course, many of the lessons learnt over the past decades will help design useful and effective AI systems. One of the central terminologies used in HCI when designing a digital system is 'usability'. Usability is defined as 'The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.' [21]. As is clear from the definition, the specification of a system's intended users, goals, and context of use play a prominent role in determining the usability of a system. Therefore, before developing any system, it is critical to understand the intended users, their goals, and the context in which the system will be used.

The tragic story of Therac-25 is an often-presented case that highlights the potential consequences of a technology that does not sufficiently consider the user or the daily use of the system in its design [26]. Therac-25 was a computerized radiation therapy system used in the USA and Canada to treat patients in the 1980s. After operators of the machine entered the desired radiation dosage, the machine sometimes gave radiation to patients that was hundreds of times greater than the norm. Following several fatal accidents, an investigation revealed that a combination of different factors, including errors in software design, development practices, and user interaction, resulted in massive overdoses of the radiation administered to patients [26]. Considerations for human factors and the design of user interfaces, the layer through which users provide input to a system and subsequently receive output, have since grown substantially. By 1995, James Reason highlighted that 'Human rather than technical failures now represent the greatest threat to complex and potentially hazardous systems.' [34]. HCI and related disciplines have since studied human error in a variety of healthcare contexts, for example in the frequently occurring task of numerical entry [11] and in clinical decision making such as determining correct prescriptions [25]. This increased understanding of human errors has resulted in numerous heuristics to support in the design of user interfaces, see *e.g.* [31, 42].

Moving beyond a focus on the usability of individual systems, the 1990s saw a shift in thinking towards supporting groups of people working together. Described by Bannon as a transition 'from human factors to human actors' [4], the interaction between users and system was no longer restricted to a one-on-one relationship but allowed for multiple persons to make use of one or even multiple devices collaboratively. Popular platforms like 'Google Docs' and 'Microsoft Teams' highlight the possibilities of this type of multi-device collaboration. Examples within the healthcare context include the early study of working collaboratively and simultaneously with digital and analogue patient records [28], the physical coordination and distribution of

work in hospitals [20], and designing technology to support telemedicine [24]. See Fitzpatrick and Ellingsen for a detailed overview of collaboration-focused contributions in the healthcare context [17].

A last major direction within HCI research is a move towards ubiquitous and context-aware computing [5]. Ubiquitous computing describes the notion that computing can appear at any location and is no longer restricted to dedicated desktop-based workplaces [41]. Instead, tablets and smartphones now enable healthcare workers to input and interact with patient data anywhere. These technical possibilities have allowed healthcare technology to leave the hospital and be part of patients' daily lives through smartphone applications and wearable devices [3]. The increasing miniaturization of hardware has furthermore enabled the use of sensors to monitor patients' vital signs or track the flow of patients and equipment in the hospital [2].

Modern digital systems often combine all these focus areas, recognizing the need for usable systems that support real-world collaboration and combine the unique capabilities of multiple devices. As such, prior work in HCI provides the building blocks for AI technology that is usable and aligns with the needs of its end-users. To make Al systems 'work' in the modern healthcare environment, system designers need to provide a clearly understood interface that supports patient safety, enables collaboration between clinical team members, and support data collection, evaluation, and decision-making both inside the hospital and outside in the patients' day-to-day life.

41.3 Designing Human-AI in Healthcare

The design of AI-driven systems introduces novel challenges. In comparison to non-AI based systems, in which the system's behaviour can be largely pre-defined, the uncertainty of the system's behaviour is much higher [8]. Furthermore, given the novelty of the technology, users are likely uncertain as to what level of support they can expect and, equally important, what type of support the system is unable to provide. Other questions include the long-term effect of the technology on skills development and the possible replacement of human workers. Like the design of non-AI based systems, however, is vitally important to involve the envisioned end-users and other relevant stakeholders in the early stages of the design process. In this section, we outline recommendations for the design of human-AI systems in a healthcare context.

41.3.1 Stakeholder Involvement

HCI researchers have long argued for the inclusion of relevant stakeholders early on in the design and development process. Within the health domain, these stakeholders typically consist of medical professionals (*e.g.*, nurses, clinicians) and other relevant roles such as patients, family members, and

hospital support staff. By involving relevant stakeholders early on in designing a new system, we can grasp the problem domain more clearly, thereby increasing the chance that the eventual outcome will align with real-world problems.

Relevant stakeholders can provide useful information to inform the system's functionality based on their prior experiences and expectations. Techniques for collecting data from these stakeholders include interviews, observations of work-related activities, and focus group discussions. Data collected via these methods are typically converted to concrete user requirements as part of the overall system development process. Different forms of end-user involvement can be considered. For an in-depth discussion of different types of end-user involvement, see *e.g.* Noyes and Baber on user-centred design [32] and Muller and Kuhn on participatory design [30]. Regardless of the exact strategy chosen for end-user involvement, involving relevant stakeholders early on in the development of medical technology is critical to ensure the applicability of the developed system in a real-world context.

Key Insight 1: Collect requirements and expectations from the intended end-users during the initial stages of the project to ensure real-world relevance and fit.

Key Insight 2: Ensure representative end-users of the system are involved in the design and evaluation of your system.

41.3.2 Explaining AI Decisions

An often-discussed downside of AI systems is the opaque nature of their behaviour, which makes it challenging to determine why certain decisions were made. Moving away from this so-called black-box behaviour, there is an increasing interest in explainable AI. Particularly in the context of healthcare, in which any decision can have far-reaching consequences, relevant stakeholders need to understand why an AI support system would recommend or even autonomously carry out a given action. Providing an insight into the reasoning of these systems can prevent errors, as a medical expert can intervene when incorrect conclusions are drawn.

What makes for a usable explanation differs between stakeholders - for example, clinicians require a different explanation model than patients. As per Key Insights 1 and 2, recognizing user needs is critical to recognizing the type of explanations that might be helpful to specific end-users. For example, prior work highlights that explanations of AI suggestions do not necessarily need to be provided immediately when generated, as such explanations can interrupt physicians during tasks such as surgery, where attention is critical [38].

Key Insight 3: Explanations of an AI system's behaviour align with the needs of the target user group.

41.3.3 Fit for Context

Existing algorithms typically form the basis of novel AI systems. While making for an effective starting point, any new system will need to be designed around the requirements of the context in which it is deployed. This includes both the technical fit of the system, such as whether the data on which the system is trained matches the real-world context in which it is deployed, as well as the fit to the user's task. A recent example of poor contextual fit is found in an AI-driven retinal assessment deployment by Google in Thailand. In analysing the system's performance during deployment, Beede et al. highlight how the system often failed when presented with the images provided by the nurses due to poor lighting conditions and the high patient throughput [7]. With more than a fifth of the images rejected, the system subsequently assigned patients to see a specialist on a different day, even when nurses assessed the image they took of a patient as negative. This highlights the potential negative consequences that a poor fit for context can have, resulting in additional work for both patients and medical professionals. When considering the fit of a system to a specific task, the interface between user and system plays a critical role. In designing the visual markers of an AI endoscopy support system, Van Berkel et al. overlaid AI suggestions on patient video footage in a colour very distinct to the colour of the colon wall to ensure that the suggestion would be most visible [38]. This design decision was subsequently evaluated with endoscopists.

Key Insight 4: Align the design of the human-AI system with the specific characteristics of the context in which it is used.

41.4 Evaluating Human-AI in Healthcare

Before deploying any AI system, it is critical to evaluate its use in a realistic setting. Although lab-based simulations can be used to indicate the accuracy of an AI system, evaluating a system in a clinical environment where people use it outside the development team can highlight previously unconsidered problems. This section outlines the key aspects in evaluating a human-AI system to be used in the healthcare context.

41.4.1 Representing Reality

When considering the evaluation of AI systems, it is critical to take into account the day-to-day reality in which the system will be used. The closer the alignment between the evaluation and reality, the more valuable the insights of a conducted evaluation are. This notion is called 'ecological validity', and is defined by Carter et al. as 'the extent to which a study comprises "real-world" use of a system' [13]. Ecological validity is not a straightforward binary state in which a study either achieves or fails to achieve ecological validity. Instead, the concept consists of a wide range of dimensions. The degree to which ecological validity can be achieved is often restricted by practical, ethical, or other limitations. Van Berkel et al. highlight seven primary dimensions of ecological validity as encountered in clinical usability evaluations: user roles, environment, training, scenario, patient involvement, software, and hardware [39]. As stressed in their article, it is impossible to obtain an 'optimal configuration' on each dimension. Instead, it is critical to identify the trade-offs presented in a study to increase ecological validity where possible. For example, while it is often unfeasible to involve actual patients in a medical evaluation, the use of simulated patients has proven effective for training and evaluation purposes [44].

Key Insight 5: Identify and describe the most relevant dimensions of an evaluation's ecological validity prior to evaluation.

Key Insight 6: Maximize a study's ecological validity within the given constraints, and actively consider the effect of these constraints on evaluation outcomes.

41.4.2 Longitudinal Assessment

Usability evaluations typically focus on questions such as 'Can a user complete the task they set out to do?' and 'Are the instructions and output results clear?'. Such usability evaluations are typically of limited duration (*i.e.*, spanning less than one hour), in which the participant is asked to complete a set of typical work-related tasks. Well suited to identifying problems in the design of user interfaces and workflow-related errors, usability evaluations cannot answer all questions in the evaluation of human-AI systems. In particular, understanding the effects of introducing a new AI system into a healthcare organization requires a longitudinal assessment.

Prior work highlights that our interaction with technology changes over time [35, 40]; with an initial interest in a new technological artifact typically waning after a while. Consider, for example, the many activity trackers collecting dust in bedroom drawers after an initial couple of weeks of intense

use. This phenomenon is known as the novelty effect. In the medical context, the non-adoption or abandonment of technological solutions is common [9, 19]. An investigation into this phenomenon by Greenhalgh et al. points to seven aspects that impact technology adoption, including the condition or illness, the technology, and the wider (institutional and societal) context [19]. Greenhalgh et al. furthermore point to the impact of time as a crucial element in an organizations' adaptation to introducing a new system. Time allows medical staff and other employees to integrate the technology into their daily workflow or opt to abandon the system altogether. Therefore, it is critical to evaluate systems over an extensive period of deployment to collect insights on the technology's adoption by the intended users.

> **Key Insight 7**: Conduct long-term evaluations to evaluate the adoption and impact of novel systems.

41.4.3 Validating across Diverse Datasets

AI-driven decision-making relies heavily on the data on which it is trained. A well-known issue in AI systems is the lack of a broad representation in these datasets, most commonly due to limitations in the diversity of race or gender in historical data. Such disparities in the data can result in unexpected and unintended consequences when the system is faced with a person from a group under-represented in the data [18]. This is far from a theoretical problem. Historical data shows that, for example, from 2003 to 2009 women were less likely to receive optimal care related to artery diseases in US hospitals as compared to men [27]. Such disparities, for example, in the rate at which patients received lipid-lowering medications, for instance, will continue with the introduction of algorithmic decision-making if that is trained on biased historical data. Furthermore, and perhaps even more concerning, is that the opaque nature of the majority of current AI systems essentially hides the underlying reasons for the decision outcome. Given the inherent risk of hidden bias in AI-driven decision-making, it is essential to evaluate an algorithm's performance across a diverse dataset and validate whether the assessments made across the dataset are valid.

> **Key Insight 8**: Validate the decisions made by an AI system across a diverse dataset to identify potential biases and shortcomings in the training data.

41.4.4 Iterative Development

Lastly, the results of any evaluation should feed back into the system's development to ensure that any identified shortcomings can be resolved. Re-

peatedly carrying out the process of development and evaluation is known as iterative development [22], and allows for a rapid and agile development process in which the initial requirements of the system are adjusted and refined based on the feedback collected during evaluations. Furthermore, it supports the development and evaluation of individual aspects of a system's software. This is often a necessity, given the many sub-systems that make up the entire system. It is generally recommended to assess the effectiveness of early-stage prototypes with the intended end-users rather than waiting for a complete system implementation before evaluating the system with users. The early evaluation of prototype implementations ultimately reduces development time, as initial end-user feedback can highlight wrong directions in the development process.

> **Key Insight 9**: Elicit feedback from end-users early on in the development process and iterate development based on user feedback.

41.5 Open Research Questions

While the development of AI has taken great leaps forward over the past decades in terms of technical capabilities, much work remains to be done to integrate and operationalize AI technology in daily medical practice. This chapter outlines recommendations and best practices for designing and evaluating AI systems in a healthcare context, while acknowledging that the development of best practices is still under development. To inspire and outline future research opportunities in this critical domain, we highlight three open research questions:

- How to support collaborative work in AI-based systems.
- How to involve patients in algorithm-driven decision-making.
- How to design effective task sharing.

41.5.1 Supporting Collaborative Work

Teamwork and collaboration play an essential role in daily healthcare practice and have been described as a necessary element to provide safe, efficient, and patient-centred care [16]. However, most AI systems in healthcare are designed around the notion of a single end-user, providing limited support and tools that allow people to share the input to an AI-support system. In a review of teamwork in dynamic healthcare contexts (*e.g.*, operating rooms, intensive care), Manser identifies patterns of communication, coordination, and leadership that positively affect clinical teamwork [29]. Future AI-driven support systems must incorporate these critical elements not only to support one clinician, but instead to be able to support a team of medical experts.

A promising direction for this work is presented in the work by Bardram and Houben on 'collaborative affordances' [6]. Collaborative affordances are defined as 'a relation between a [physical and/or digital] artifact and a set of human actors, that affords the opportunity for these actors to perform a collaborative action within a specific social context' [6]. Within the context of medical records, Bardram and Houben identify four collaborative affordances - 'portability', 'collocated access', 'shared overview', and 'mutual awareness' - which allow for collaboration between medical staff. These affordances, among others, are largely missing from contemporary medical AI systems, prohibiting information sharing and collaboration.

41.5.2 Shared Decision-Making

Effective communication between patients and medical professionals is critical in achieving a successful care trajectory [15]. A widely established concept within the patient-professional relationship is that of shared decision-making [14, 15]. Charles et al. define shared decision-making via four criteria: '(1) that at least two participants, physician and patient, be involved; (2) that both parties share information; (3) that both parties take steps to build a consensus about the preferred treatment; and (4) that an agreement is reached on the treatment to implement.' [14]. Elwyn et al. highlight that shared decision-making is based on the introduction of choice to the patient, describing different treatment options, and assisting patients in exploring their preferences to arrive at a decision [15]. As seen from these definitions, it is critical for a patient to understand the benefits, trade-offs, and necessity of different treatment options.

The introduction of AI-support systems essentially introduces a third party to the shared decision-making process, in which an AI system can support in the identification and selection of suitable treatment options. While the concept of AI explainability has been explored from the perspective of medical professionals (see Cai et al. on dealing with imperfect AI in pathology [10]), the patient perspective in dealing with AI remains underexplored in the current literature. Determining where and how AI support can be integrated into the shared decision-making process is an essential research trajectory given the increased reliance on AI technology by clinical professionals. Failing to ensure that patients can understand and engage with this new technology will ultimately undermine the concept of shared decision-making.

41.5.3 Effective Task Sharing

The ongoing development of AI will increase the types of tasks that AI systems can carry out autonomously. This makes AI support systems more powerful and will require more effective techniques for sharing (sub-)tasks

between the human operator and the AI system. For example, even advanced AI systems will encounter scenarios in which their uncertainty exceeds a pre-determined threshold. At this point, a human expert must step in and take over from the AI system to make a decision. This behaviour can already be found in today's autonomous driving systems and will increasingly manifest itself in the healthcare context.

Sujan et al. highlight the various steps required for such task handovers to be successful: the AI system has to recognize its limitations, decide on which elements to hand over, determine how this handover should be accomplished, and decide when it ought to be carried out [36]. Each of these steps requires further investigation of user needs, contextual requirements, and technical capabilities. Prior work has suggested the design of structured communication protocols between the user and AI system, as currently exist in handover protocols between medical professionals [36]. Similarly, a medical professional handing over a task to an AI system requires a pre-defined protocol and interface that outline the expected AI behaviour and provide appropriate mechanisms for the user to intervene if needed.

41.6 Conclusion

Despite the numerous challenges facing the deployment of AI in real medical contexts and applications, AI is likely to transform modern-day healthcare. To maximize its potential, however, we need to ensure that the AI systems we put out into the world align with the needs of medical professionals, patients, and other relevant stakeholders. This chapter outlines the primary steps required to design and evaluate AI systems that bridge computational possibilities and real-world needs and conditions. We highlight the necessity of involving relevant stakeholders early on in the design of human-AI systems. Insights obtained from these stakeholders inform the intended end-users' needs, for example in critical areas such as AI explainability and the unique aspects of the context of use. Subsequently, this chapter provides recommendations for evaluating human-AI systems in a healthcare context, stressing the need for ecologically valid and longitudinal evaluations to obtain the most relevant insights. AI-driven systems provide novel challenges in their evaluation due to the relative unpredictability of their behaviour and concerns about bias introduced into AI decision-making. Furthermore, we highlight open research questions, indicating that while AI technology has made tremendous leaps forward, an extensive amount of work is still required to provide medical professionals with AI systems that can truly be embedded into their everyday tasks. We hope that the recommendations outlined in this work can provide a stepping stone for both medical professionals and system developers who are new to the application of human-AI in healthcare.

References

[1] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. *Guidelines for Human-AI Interaction*, page 1–13. Association for Computing Machinery, New York, NY, USA, 2019.

[2] Rodolfo S. Antunes, Lucas A. Seewald, Vinicius F. Rodrigues, Cristiano A. Da Costa, Luiz Gonzaga Jr., Rodrigo R. Righi, Andreas Maier, Björn Eskofier, Malte Ollenschläger, Farzad Naderi, Rebecca Fahrig, Sebastian Bauer, Sigrun Klein, and Gelson Campanatti. A survey of sensors in healthcare workflow monitoring. *ACM Comput. Surv.*, 51(2), April 2018.

[3] Stinne Aaløkke Ballegaard, Thomas Riisgaard Hansen, and Morten Kyng. Healthcare in everyday life: Designing healthcare services for daily life. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, page 1807–1816, New York, NY, USA, 2008. Association for Computing Machinery.

[4] Liam J. Bannon. From human factors to human actors: The role of psychology and human-computer interaction studies in system design. In Ronald M. Baecker, Jonathan Grudin, William A.S. Buxton, and Saul Greenberg, editors, *Readings in Human–Computer Interaction*, Interactive Technologies, pages 205–214. Morgan Kaufmann, 1995.

[5] Jakob E. Bardram. Applications of context-aware computing in hospital work: Examples and design principles. In *Proceedings of the 2004 ACM Symposium on Applied Computing*, SAC '04, page 1574–1579, New York, NY, USA, 2004. Association for Computing Machinery.

[6] Jakob E. Bardram and Steven Houben. Collaborative affordances of medical records. *Computer Supported Cooperative Work (CSCW)*, 27(1):1–36, Feb 2018.

[7] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M. Vardoulakis. *A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy*, page 1–12. Association for Computing Machinery, New York, NY, USA, 2020.

[8] Jesse Josua Benjamin, Arne Berger, Nick Merrill, and James Pierce. 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, New York, NY, USA, 2021. Association for Computing Machinery.

[9] Ann Blandford. HCI for health and wellbeing: Challenges and opportunities. *International Journal of Human-Computer Studies*, 131:41–51, 2019.

[10] 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. *Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making*, page 1–14. Association for Computing Machinery, New York, NY, USA, 2019.

[11] Paul Cairns, Pratyush Pandab, and Christopher Power. The influence of emotion on number entry errors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, page 2293–2296, New York, NY, USA, 2014. Association for Computing Machinery.

[12] J. M. Carroll. Human-computer interaction: psychology as a science of design. *Annu Rev Psychol*, 48:61–83, 1997.

[13] Scott Carter, Jennifer Mankoff, Scott R. Klemmer, and Tara Matthews. Exiting the cleanroom: On ecological validity and ubiquitous computing. *Human–Computer Interaction*, 23(1):47–99, 2008.

[14] Cathy Charles, Amiram Gafni, and Tim Whelan. Shared decision-making in the medical encounter: What does it mean? (or it takes at least two to tango). *Social Science & Medicine*, 44(5):681–692, 1997.

[15] 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. Shared decision making: a model for clinical practice. *Journal of general internal medicine*, 27(10):1361–1367, 2012.

[16] Rachael Finn, Mark Learmonth, and Patrick Reedy. Some unintended effects of teamwork in healthcare. *Social Science & Medicine*, 70(8):1148–1154, 2010.

[17] Geraldine Fitzpatrick and Gunnar Ellingsen. A review of 25 years of CSCW research in healthcare: Contributions, challenges and future agendas. *Computer Supported Cooperative Work (CSCW)*, 22(4):609–665, Aug 2013.

[18] Milena A. Gianfrancesco, Suzanne Tamang, Jinoos Yazdany, and Gabriela Schmajuk. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. *JAMA Internal Medicine*, 178(11):1544–1547, 11 2018.

[19] Trisha Greenhalgh, Joseph Wherton, Chrysanthi Papoutsi, Jennifer Lynch, Gemma Hughes, Christine A'Court, Susan Hinder, Nick Fahy, Rob Procter, and Sara Shaw. Beyond adoption: A new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. *J Med Internet Res*, 19(11):e367, 2017.

[20] Ole Hanseth and Nina Lundberg. Designing work oriented infrastructures. *Computer Supported Cooperative Work (CSCW)*, 10(3):347–372, 2001.

[21] ISO. Ergonomics of human-system interaction. 2018.

[22] Pankaj Jalote, Aveejeet Palit, Priya Kurien, and V.T. Peethamber. Timeboxing: a process model for iterative software development. *Journal of Systems and Software*, 70(1):117–127, 2004.

[23] Bonnie Kaplan and Kimberly D. Harris-Salamone. Health IT success and failure: recommendations from literature and an AMIA workshop. *Journal of the American Medical Informatics Association : JAMIA*, 16(3):291–299, 2009.

[24] Simon M. Kaplan and Geraldine Fitzpatrick. Designing support for remote intensive-care telehealth using the locales framework. In *Proceedings of the 2nd Conference on Designing Interactive Systems: Processes, Practices, Methods,*

and Techniques, DIS '97, page 173–184, New York, NY, USA, 1997. Association for Computing Machinery.

[25] G. J. Kuperman, A. Bobb, T. H. Payne, A. J. Avery, T. K. Gandhi, G. Burns, D. C. Classen, and D. W. Bates. Medication-related clinical decision support in computerized provider order entry systems: a review. *J Am Med Inform Assoc*, 14(1):29–40, 2007.

[26] N.G. Leveson and C.S. Turner. An investigation of the therac-25 accidents. *Computer*, 26(7):18–41, 1993.

[27] Shanshan Li, Gregg C. Fonarow, Kenneth J. Mukamal, Li Liang, Phillip J. Schulte, Eric E. Smith, Adam DeVore, Adrian F. Hernandez, Eric D. Peterson, and Deepak L. Bhatt. Sex and race/ethnicity-related disparities in care and outcomes after hospitalization for coronary artery disease among older adults. *Circulation: Cardiovascular Quality and Outcomes*, 9:S36–S44, 2016.

[28] Paul Luff, Christian Heath, and David Greatbatch. Tasks-in-interaction: Paper and screen based documentation in collaborative activity. In *Proceedings of the 1992 ACM Conference on Computer-Supported Cooperative Work*, CSCW '92, page 163–170, New York, NY, USA, 1992. Association for Computing Machinery.

[29] T. Manswer. Teamwork and patient safety in dynamic domains of healthcare: a review of the literature. *Acta Anaesthesiologica Scandinavica*, 53(2):143–151, 2009.

[30] Michael J. Muller and Sarah Kuhn. Participatory design. *Commun. ACM*, 36(6):24–28, June 1993.

[31] Jakob Nielsen. Enhancing the explanatory power of usability heuristics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '94, page 152–158, New York, NY, USA, 1994. Association for Computing Machinery.

[32] Jan Noyes and Chris Baber. *User-centred design of systems*. Springer Science & Business Media, 1999.

[33] Raj M Ratwani, Erica Savage, Amy Will, Ryan Arnold, Saif Khairat, Kristen Miller, Rollin J Fairbanks, Michael Hodgkins, and A Zachary Hettinger. A usability and safety analysis of electronic health records: a multi-center study. *Journal of the American Medical Informatics Association*, 25(9):1197–1201, 07 2018.

[34] J Reason. Understanding adverse events: human factors. *BMJ Quality & Safety*, 4(2):80–89, 1995.

[35] Sofia Serholt and Wolmet Barendregt. Robots tutoring children: Longitudinal evaluation of social engagement in child-robot interaction. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction*, NordiCHI '16, New York, NY, USA, 2016. Association for Computing Machinery.

[36] Mark Sujan, Dominic Furniss, Kath Grundy, Howard Grundy, David Nelson, Matthew Elliott, Sean White, Ibrahim Habli, and Nick Reynolds. Human factors challenges for the safe use of artificial intelligence in patient care. *BMJ Health & Care Informatics*, 26(1):e100081, 2019.

[37] The Turing Institute. Data science and AI in the age of COVID-19. 2021.

[38] Niels van Berkel, Omer F. Ahmad, Danail Stoyanov, Laurence Lovat, and Ann Blandford. Designing visual markers for continuous artificial intelligence support: A colonoscopy case study. *ACM Trans. Comput. Healthcare*, 2(1), December 2021.

[39] Niels van Berkel, Matthew J. Clarkson, Guofang Xiao, Eren Dursun, Moustafa Allam, Brian R. Davidson, and Ann Blandford. Dimensions of ecological validity for usability evaluations in clinical settings. *Journal of Biomedical Informatics*, 110:103553, 2020.

[40] Niels van Berkel, Simon Dennis, Michael Zyphur, Jinjing Li, Andrew Heathcote, and Vassilis Kostakos. Modeling interaction as a complex system. *Human–Computer Interaction*, 36(4):279–305, 2021.

[41] Mark Weiser. The computer for the 21 st century. *Scientific American*, 265(3):94–105, 1991.

[42] Sarah Wiseman, Anna L. Cox, and Duncan P. Brumby. Designing devices with the task in mind: Which numbers are really used in hospitals? *Human Factors*, 55(1):61–74, 2013.

[43] Wei Xu. Toward human-centered AI: A perspective from Human-Computer Interaction. *Interactions*, 26(4):42–46, 2019.

[44] Amitai Ziv, Paul Root Wolpe, Stephen D. Small, and Shimon Glick. Simulation-based medical education: An ethical imperative. *Academic Medicine*, 78(8), 2003.