Sketching AI Concepts with Capabilities and Examples: AI Innovation in the Intensive Care Unit

Nur Yildirim
Carnegie Mellon University
Pittsburgh, PA, USA
yildirim@cmu.edu

Susanna Zlotnikov
Carnegie Mellon University
Pittsburgh, PA, USA
susanna.zlotnikov@gmail.com

Deniz Sayar
Izmir University of Economics
Izmir, Turkey
sayardeniz@gmail.com

Jeremy M. Kahn
University of Pittsburgh
Pittsburgh, PA, USA
jeremykahn@pitt.edu

Leigh A. Bukowski
University of Pittsburgh
Pittsburgh, PA, USA
lab108@pitt.edu

Sher Shah Amin
University of Pittsburgh
Pittsburgh, PA, USA
amin.shershah@gmail.com

Kathryn A. Riman
University of Pittsburgh
Pittsburgh, PA, USA
kathrynriman@pitt.edu

Billie S. Davis
University of Pittsburgh
Pittsburgh, PA, USA
bid8@pitt.edu

John S. Minturn
University of Pittsburgh
Pittsburgh, PA, USA
jsm120@pitt.edu

Andrew J. King
University of Pittsburgh
Pittsburgh, PA, USA
andrew.king@pitt.edu

Dan Ricketts
University of Pittsburgh
Pittsburgh, PA, USA
d.r@pitt.edu

Lu Tang
University of Pittsburgh
Pittsburgh, PA, USA
lutang@pitt.edu

Venkatesh Sivaraman
Carnegie Mellon University
Pittsburgh, PA, USA
vsivaram@andrew.cmu.edu

Adam Perer
Carnegie Mellon University
Pittsburgh, PA, USA
adamperer@cmu.edu

Sarah M. Preum
Dartmouth College
Hanover, NH, USA
sarah.masud.preum@dartmouth.edu

James McCann
Carnegie Mellon University
Pittsburgh, PA, USA
jmccann@cs.cmu.edu

John Zimmerman
Carnegie Mellon University
Pittsburgh, PA, USA
johnz@cs.cmu.edu

ABSTRACT

Advances in artificial intelligence (AI) have enabled unprecedented capabilities, yet innovation teams struggle when envisioning AI concepts. Data science teams think of innovations users do not want, while domain experts think of innovations that cannot be built. A lack of effective ideation seems to be a breakdown point. How might multidisciplinary teams identify buildable and desirable use cases? This paper presents a first hand account of ideating AI concepts to improve critical care medicine. As a team of data scientists, clinicians, and HCI researchers, we conducted a series of design workshops to explore more effective approaches to AI concept ideation and problem formulation. We detail our process, the challenges we encountered, and practices and artifacts that proved effective. We discuss the research implications for improved collaboration and stakeholder engagement, and discuss the role HCI might play in reducing the high failure rate experienced in AI innovation.

CCS CONCEPTS

• **Human-centered computing** → **Interaction design process and methods**.

KEYWORDS

Brainstorming, ideation, human-centered AI, healthcare

ACM Reference Format:
Nur Yildirim, Susanna Zlotnikov, Deniz Sayar, Jeremy M. Kahn, Leigh A. Bukowski, Sher Shah Amin, Kathryn A. Riman, Billie S. Davis, John S. Minturn, Andrew J. King, Dan Ricketts, Lu Tang, Venkatesh Sivaraman, Adam Perer, Sarah M. Preum, James McCann, and John Zimmerman. 2024. Sketching AI Concepts with Capabilities and Examples: AI Innovation in the Intensive Care Unit. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3613904.3641896

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3641896

1 INTRODUCTION

Artificial intelligence (AI) is transforming the landscape of healthcare. From cancer diagnosis [19] to prognosis [136], automated documentation [71], and treatment recommendations [109], AI applications in healthcare offer the promise of improved clinician

experience and better healthcare outcomes for patients. While AI's technical advances showcase impressive performance in lab settings, AI systems largely fail when moving to clinical practice [42, 92, 120, 137, 145]. HCI researchers note that the clinical utility and actionability –whether clinicians can take specific actions based on a prediction– of healthcare AI applications often remain unclear [44, 119, 142]. Clinicians do not use AI systems, often because systems do not deliver what they need.

The challenge of making AI advances useful in real world contexts is not unique to healthcare. Today, the majority of AI initiatives fail, as they fail to generate enough value for users or for service providers [38, 63, 126]. Product teams share experiencing repeated AI failures due to selecting and working on the wrong problem – high-risk projects that may or may not be valuable or that entail unavoidable challenges around fairness and bias [13, 54, 95, 140]. AI development practices remain technology-driven with little attention to human needs and wants [140]. Stakeholders that do not have a background in data science or AI are rarely involved in conversations around the objective of the underlying model or the overall problem formulation, if involved at all [29, 39]. Challenges in multidisciplinary collaboration across team members poses a major barrier to AI design and development [74, 88, 95, 96]. As AI capabilities become readily available, a critical question arises: How can multidisciplinary innovation teams effectively identify low-risk, high-value AI use cases?

In response to these challenges, HCI researchers called for human-centered, participatory approaches to AI development –especially in early ideation and problem formulation phases– to reduce the risk of developing unwanted technology [29, 135, 140]. Studies on industry best practices revealed that effective innovation teams brainstorm using AI capabilities and examples of AI applications to close knowledge gaps between data science, HCI, and domain expertise [134, 138, 140]. These emergent AI innovation practices resembled a blend of user-centric and tech-centric approaches, where teams rapidly generated many AI concepts to match [10] AI capabilities with human needs within a specific product domain [140]. An emerging body of research have started to explore how team members and domain stakeholders might envision and co-design AI use cases (e.g., in law [27], public services [112], accessibility [122]), and the types of design process processes, tools, and methods that might prove effective [34, 75, 87].

Building on this line of research, we set out to explore clinically relevant and feasible AI uses cases for intensive care within a multidisciplinary team of AI researchers, HCI researchers, data scientists, and healthcare professionals. Our prior work detailed the development of *the AI Brainstorming Kit* [139] – a resource to help HCI experts facilitate AI concept ideation within multidisciplinary teams, especially to identify low-risk, high-value concepts where *moderate AI performance* can create value. In this paper, we present a reflective account of our design process as a case study of early phase AI innovation, with a specific focus on capturing the iterative research activities. Our team had access to a rich ICU dataset (similar to MIMIC [62]) that was collected across 39 intensive care units (ICUs) from 18 hospitals. We engaged in an iterative design process to broadly explore the problem-opportunity space for *getting the right design* [17]. We conducted a three-phase study, where we moved from ideation to problem formulation, concept design, and initial assessment with end users.

Phase 1 Brainstorming focused on envisioning many AI concepts and use cases for the ICU, before selecting and building an application. We conducted two brainstorming workshops within our multidisciplinary team. The first workshop followed a traditional user-centered design approach with a focus on user needs. The second workshop combined user-centered and matchmaking [10] approaches to consider both user needs *and* AI capabilities simultaneously. Building on *the AI Brainstorming Kit [139]*, we used a set of AI capabilities and examples to scaffold ideation, selecting examples where moderate model performance was 'good enough' to produce value. An assessment of outputs from each workshop demonstrated that the latter approach resulted in more effective brainstorming with many concepts that were low-risk in terms of feasibility and clinician acceptance, and medium to high-value for clinicians.

Phase 2 Problem Formulation focused on detailing a subset of AI concepts further *(e.g., predicting medication availability and anticipatory ordering)*. Our brainstorming sessions yielded many concepts that leveraged AI capabilities in ways that provided utility for clinicians; however, it was unclear whether and how these could operationalize our unique ICU dataset. To tackle this challenge, we conducted a follow-up workshop session, where we detailed the required model performance, point of interaction, data requirements, and risks (e.g., consequences of potential errors) for 12 use cases we previously identified. We created a worksheet detailing the data, model reasoning, and interaction form to disentangle interaction design and model building considerations. This proved an effective artifact for refining concepts, revealing unreliable data, and considering if simpler versions of a concept might also be valuable.

Phase 3 Sketching and Co-Design explored further refining an AI concept towards prototyping to elicit early phase user feedback. We selected a concept that aimed to predict if a patient is eligible to receive the protocol for assessing readiness for liberation from mechanical ventilation. We created sketches detailing the concept. We conducted four co-design workshops with 11 clinicians to probe whether and how this concept might support them in considering and executing this specific evidence-based protocol. Participants perceived the concept as valuable, and they articulated detailed design requirements for interaction design as well as model building and data.

This paper makes two contributions. First, we present a rare case study of early phase AI innovation within a multidisciplinary team of data scientists, domain experts, and HCI researchers. We describe the challenges faced across brainstorming, concept design, and initial assessment. These practices serve as a starting point for multidisciplinary teams to structure design activities for navigating early phase human-centered AI innovation. Second, we discuss remaining challenges and outline opportunities for HCI researchers to better support and facilitate effective collaboration and stakeholder engagement in AI innovation projects, specifically to identify low-risk, high-value use cases in high-stakes contexts including healthcare and beyond.

2 RELATED WORK

2.1 Challenges of AI Product Innovation

HCI research characterizes *technology as a material* [102, 128] that designers can explore to envision novel interactions (e.g., bluetooth [115], haptics [85], and software [93]). Embracing this material lens, researchers have framed 'AI as design material' to explore AI's opportunities and challenges for HCI research and practice [114, 135, 138]. From language translation to text summarization, medical diagnosis and image generation, technical AI advances offer unprecedented capabilities. While these advances open up a novel space for interactive systems, they also pose unique challenges to designing AI products and services [135]. A large body of research investigated the challenges around explainability [1, 77], trust and reliance [15], user control [108], feedback [113], error recovery [73], and fairness-related harms [124], just to name a few.

While research efforts have largely focused on mitigating issues that arise post-deployment, recent research points to a more consequential problem: more than 85% of AI innovation projects fail pre-deployment [38, 63, 126]. Failure includes taking on projects that are too complex or infeasible; selecting problems that entail unavoidable fairness issues, such as privacy concerns or algorithmic bias; and building systems that in the end fail to generate enough value for customers or service providers [126]. Some researchers critiqued this breakdown from a perspective of 'validity', raising the importance of asking whether an AI system provides any benefits in the first place [98]. Studies investigating industry practices attribute AI failures to lack of human-centered approaches and ineffective collaboration between cross-disciplinary team members in early problem formulation phases of a project [31, 74, 81, 88, 96, 123, 125, 140].

In recent years, resources in the form of guidelines and toolkits became available to address some of AI's design challenges (e.g., human-AI guidelines [2, 3, 94], fairness toolkits [30]). However, investigations on how teams use these resources indicate that these resources mainly help at later stages, *after* problem selection and formulation. Practitioners ask for resources that support early phase ideation and problem formulation to discover use cases where AI might be a good solution [140]. A related strand of research investigating industry best practices revealed that effective innovation teams work with AI capabilities and examples to scaffold cross-disciplinary ideation [134, 138, 140, 151]. These resources detail what AI can do instead of how AI works using non-technical terms (e.g., detect customer patterns; predict seasonality trends), which seem to help user experience designers and product managers gain a practical understanding of AI [140].

Finally, researchers report limitations of user-centered design (UCD) in AI innovation [41, 46, 91, 131, 140, 148], and highlight emergent design processes that blend UCD and *matchmaking* [10] – an innovation process that starts with a technical capability to systematically search for customers that might benefit from it. Researchers also point out that innovation teams often focus on complex use cases where near-perfect AI performance is needed for a concept to be useful [40, 135]. A recent analysis of 40 AI applications note that the majority of real-world applications in fact leverage moderate model performance, suggesting that teams should focus on cases where imperfect AI can create value [139].

HCI researchers have explored AI concept ideation through design-led inquiry to provide first-person accounts of their design process, challenges, and emerging solutions [7, 69, 70, 132, 152]. For example, Yang et al. detailed how a team of HCI and NLP researchers envisioned and prototyped AI-powered features for Microsoft Word [132]. Kayacik et al. described how UX designers and AI research scientists envisioned AI-driven concepts using generative AI capabilities for music creation [69]. In the same spirit, we set out to contribute a detailed case study of our ideation process for envisioning and designing AI use cases for intensive care.

2.2 Broadening Participation in AI Design

A growing body of work has called for socio-technical, participatory approaches to meaningfully engage domain stakeholders throughout the AI development lifecycle [6, 24, 28, 29, 117, 149]. Prior research notes that stakeholders with little to no background in data science or AI are rarely involved in problem selection and formulation, if involved at all [29, 39, 55, 67]. There is a knowledge gap between data science and domain expertise [74, 132, 140]: Domain experts and designers struggle to understand what AI can do, they often envision AI services that cannot be built [35, 78, 135, 144]. Data scientists find it challenging to elicit needs from domain experts, and without this input, they tend to envision AI services that users and impacted stakeholders do not want [74, 81, 88, 96]. Teams do not seem to ideate; they focus on building a single application without exploring the space of possibilities [140].

Recent HCI research has proposed new design methods, artifacts, and resources, such as metaphors [34, 87], AI lifecycle comicboarding [75], onboarding materials [20], and other artifacts [4, 76] to facilitate effective stakeholder engagement. Notably, research employing this type of resources often focuses on later stage AI phases, detailing how to refine existing AI systems or mitigate harmful outputs. Relatively little research has offered a detailed account of early phase ideation and problem formulation with domain experts and impacted stakeholders. Few examples worth noting present case studies on envisioning and designing AI use cases in child welfare [112], fact-checking [80], law [27], and content moderation [50], and accessibility [84, 122]. We draw on this strand of research to explore effective design processes and activities for engaging clinical domain stakeholders in AI concept ideation. Specifically, we utilize a design ideation resource, namely *the AI Brainstorming Kit* [139], that we developed in our prior work to explore *how* to navigate early phase AI innovation within a multidisciplinary team.

2.3 Designing AI for Healthcare

Healthcare is a complex product-service ecosystem consisting of many stakeholders (e.g., clinicians, patients, healthcare managers, insurance providers, regulators, etc) [72, 90, 121]. A large body of research has explored the iterative design of healthcare products and services with a focus on stakeholder engagement in the early design stages [8, 26, 57, 99, 104, 141, 147]. In recent years, the advances in AI and the availability of high-density datasets, such as patient electronic health records (EHR), have enabled a new wave of innovations, spanning systems that support diagnosis, treatment recommendations, and automated documentation. However, similar to other domains, AI systems in healthcare have a poor track record; they largely fail when moving from research labs to clinical practice

[25, 32, 61, 107, 127, 137, 142]. The clinical utility of these systems remain often unclear [42, 44, 120]; as a result, clinicians often do not use them [119, 137]. Recent HCI research has developed healthcare AI systems with special attention to challenges around workflow integration [5, 16], calibrating clinician trust [59, 109, 133], transparency and setting mental models [19, 53], and risks of biases and harm [129]. Relatively little work engaged healthcare stakeholders in the early stages of AI development to envision concepts that leverage AI capabilities or explore data requirements with an eye for downstream applications [90, 133]. Our work aims to address this gap, specifically within the context of intensive care.

The intensive care unit (ICU) is a complex, team-based healthcare setting involving many clinicians (e.g., attending physicians, fellows, residents, nurse practitioners, respiratory therapists) providing round-the-clock care for critically ill patients [100]. Prior HCI work on ICUs focused on conducting field studies to understand clinician needs and workflows [65, 100, 101, 143], and developing technical systems and interventions (e.g., automating patient note documentation [48, 130], reducing alert fatigue and interruptions [18, 23, 111]). AI research advances in ICU demonstrate systems that predict treatment medications [116], predict if a patient will need a ventilator [116], predict patient discharge and readmission [79], and predict the onset of conditions like sepsis [89]. While these proof-of-concept models indicate an initial feasibility, it remains unclear whether clinicians need help with these tasks. A recent study interviewed ICU physicians and nurses to elicit *what predictions would be useful* [37] and found that clinicians desire predictions around patient trajectory and prioritization, mainly to reduce the high cognitive load rather than help with decision making. We build on this line of work to explore data and AI as design materials for ICU to identify clinically relevant and feasible AI use cases.

3 OVERVIEW OF DESIGN PROCESS

We wanted to develop more effective approaches to multidisciplinary brainstorming of AI concepts, especially in the early phases of ideation and problem formulation. Building on prior literature that noted successful AI innovation teams ideate before selecting what to build, we set out to tackle the challenge of ideation within a project that focused on AI innovation in the ICU.

Our academic research team (n=22) included 6 HCI, 6 data science, and 10 healthcare experts. The HCI researchers had backgrounds in interaction design, service design, and data visualization; they brought expertise in human-AI interaction and ideation. The data science members had backgrounds in data analytics, healthcare analytics, and AI research; they brought expertise in AI capabilities and what could be built with the dataset. The healthcare members all had experience in critical care medicine and included 4 attending physicians, 2 fellows, 2 nurses, and 2 non-clinical healthcare experts. They brought expertise in clinician needs. Table 1 provides a summary of our teams' composition.

We engaged in an iterative, reflective design process [17, 103, 150] to explore AI opportunities for the ICU, particularly to search for use cases that leveraged our ICU dataset. We conducted a three-phase study. The first phase focused on brainstorming; we conducted two ideation workshops within our team to identify clinically relevant and buildable use cases. The second phase focused on problem formulation; we conducted a design workshop to detail a subset of 12 concepts. The third phase focused on sketching and co-design; we created low fidelity sketches for an AI concept we had generated. We conducted four co-design sessions with 11 clinicians who had not been involved in our study to elicit feedback on the design concept. Below, we provide a brief overview of the ICU dataset our team had access to. We then present each phase in subsequent sections, unpacking the research goals, design activities, and insights gained.

Table 1: Our team consists of data science and AI researchers (DS), clinicians and healthcare experts (H), and human-computer interaction researchers (HCD).

ID	W1	W2	W3	Role	Exp.	Gn.
DS1	✓		✓	Data Scientist	10+yrs	F
DS2	✓	✓	✓	Data Scientist	3-5 yrs	M
DS3	✓	✓	✓	Data Analyst	5-7 yrs	M
DS4	✓	✓		Healthcare Analyst	10+yrs	M
DS5	✓		✓	AI Researcher	5-7 yrs	M
DS6	✓			AI Researcher	5-7 yrs	F
H1	✓	✓		ICU Physician	10+ yrs	M
H2			✓	ICU Physician	10+yrs	F
H3		✓	✓	ICU Physician	10+yrs	F
H4	✓	✓	✓	ICU Physician	10+yrs	M
H5		✓		Critical Care Fellow	5-7 yrs	F
H6		✓		Critical Care Fellow	5-7 yrs	M
H7	✓	✓	✓	Nurse Practitioner	5-7 yrs	F
H8		✓	✓	Nurse Practitioner	5-7 yrs	F
H9	✓	✓	✓	Healthcare expert	10+ yrs	F
H10	✓			Healthcare expert	10+ yrs	M
HCD1	✓		✓	HCI/AI Researcher	10+yrs	M
HCD2	✓	✓		HCI/AI Researcher	3-5 yrs	M
HCD3	✓	✓	✓	HCI Researcher	10+yrs	M
HCD4	✓	✓	✓	HCI Researcher	5-7 yrs	F
HCD5	✓	✓	✓	Service designer	5-7 yrs	F
HCD6	✓	✓	✓	Service designer	5-7 yrs	F

3.1 The ICU Dataset

The objective of our project was to broadly explore how our ICU dataset might be used to improve critical care medicine. Data availability is crucial for enabling AI capabilities [135]. However, prior studies on envisioning future AI solutions often do not draw from a particular dataset, and instead focus on what would be possible with pretrained models or data that could be collected [78, 84, 132]. While there is research exploring real-world datasets with domain experts, these studies often do not focus on AI innovation or technical feasibility of the envisioned systems [11, 36]. Our focus was on bounding ideation with a real world data set to address this gap.

Our dataset consisted of two parts: electronic healthcare records (EHR) and staffing metadata. Similar to the publicly available MIMIC dataset [62], the EHR data included patient level variables, such as hospitalization (e.g., age, gender, race, discharge disposition, admission and discharge dates, etc.); diagnosis and procedure codes, comorbidities; medications; clinical events, mechanical ventilation;

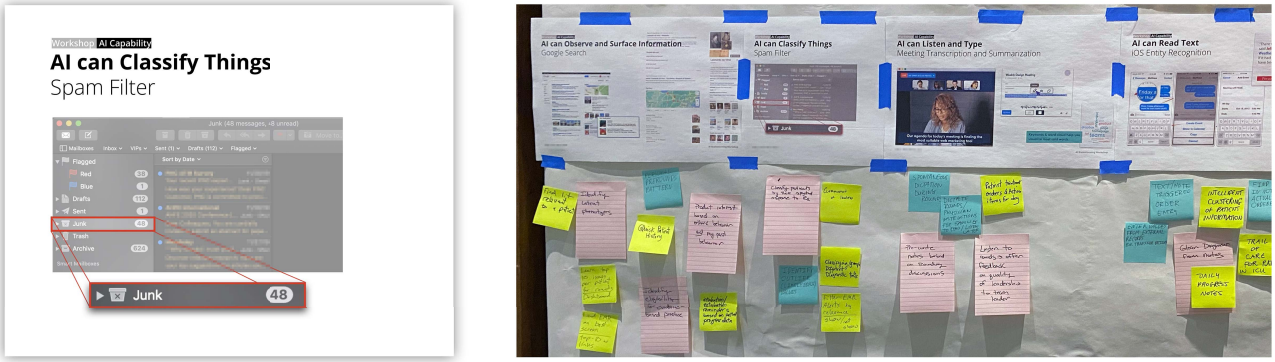

Figure 1: An AI capability abstraction and example (left), poster printouts to prompt ideation across each capability (right).

and others with a total of 15 variables. The staffing metadata included the transformation of patient level variables to anonymously identify the unique care providers across different roles (i.e., physicians, nurses, respiratory therapists) who provided primary care for each patient at a shift-level. The creation of this additional dataset was motivated by prior literature that indicated whether and how long individual care providers had worked together in the same team impacts the quality of care in the ICU [33]. The dataset was collected across 39 ICUs from 18 hospitals on the East Coast of the United States between 2018 and 2020 (see supplementary materials for a high-level overview of the data schema).

4 PHASE 1: BRAINSTORMING AI CONCEPTS

We wanted to explore how we can effectively brainstorm AI concepts as a multidisciplinary team. The healthcare members would bring expertise on what is relevant and what might transform critical care practice. The data science members would bring expertise in what might be possible to build. The HCI members would bring expertise in ideation. Our goal was to rapidly and broadly explore the problem-solution space to identify clinically relevant and buildable AI concepts to improve intensive care medicine. Our prior research presented an overview of this initial phase in the context of the development and assessment of *the AI Brainstorming Kit* – a resource that captured AI capabilities and real-world examples to scaffold cross-disciplinary ideation [139]. In this work, we provide a detailed account of the methodology and elaborate on workshop facilitation, selection of AI examples, and concept assessment and prioritization.

4.1 Method

We chose to conduct design workshops, a commonly used method in design-driven innovation [36, 103]. We conducted two workshops within our team. Each workshop had 15-17 participants involving at least one participant from each role (i.e., physician, nurse, healthcare expert, data scientist, HCI researcher). Table 1 details the involvement of participants in each workshop session. Workshops were sequential such that the outcome of a workshop informed the goals and activities of the following workshop.

A part of the challenge was the preparation and structuring of the brainstorming activities. Below, we present our thinking behind each workshop, along with details on the set of activities.

4.1.1 Workshop 1: User-centered Approach. Our first workshop followed a traditional user-centered approach. In preparation for the workshop, we had informal discussions to elicit the domain expertise of our healthcare team members. We discussed pain points and potential themes for brainstorming, both based on lived experiences and our expertise working in healthcare innovation. These preparations resulted in "how might we" prompts that we used to drive ideation (e.g., *How might we help clinicians in orchestrating a sequence of tasks? How might we support the adoption of evidence-based practice? How might we reduce clinicians' burden with documentation tasks?*). Inspired by design thinking methods [58], we set our objectives as 'thinking outside the box' and 'deferring judgment' to let go of thinking about the limits of technology.

We conducted a 2-hour in-person workshop. The workshop agenda included the introduction of goals (10 min), two consecutive ideation sessions with a short break in between (30 min), impact-effort assessment of concepts (30 min), and a short debriefing and reflection (10 min). During the ideation sessions, each team member reviewed the how-might-we prompts to first ideate individually. They next shared concepts within the group to brainstorm collectively. We used large papers, sticky notes, and markers to note down concepts. At the end of the session, we selected a subset of concepts based on the team's interest, and placed these on a large impact-effort matrix [49] by getting group consensus on whether the concept was relevant and useful to critical care (impact) and if it would be easy or difficult to implement (effort). Following the workshop, the lead HCI researcher further analyzed concepts to assess the coverage of design space (see section 4.1.3).

4.1.2 Workshop 2: User-centered and Tech-centered Approach. Following the first workshop, we had concerns that our concepts mostly focused on places where near-perfect AI performance was needed for the use cases to be valuable – a well-documented pitfall in AI design literature [35, 40, 114, 135]. Building on recent research [140], we decided to bring elements from the matchmaking method [10] to blend user-centered thinking and tech capability-driven

approaches. Prior to the workshop, we selected a subset of AI capabilities and examples from the AI Brainstorming Kit [139]. Hoping to move away from envisioning use cases that required high AI accuracy or performance, we mostly selected examples where moderate performance and imperfect AI capabilities produced value.

The capabilities and examples included *observe and surface information* (contextual web search); *classify things* (spam filter); *listen and type* (real-time meeting transcription); *read text* (text message entity recognition); *predict text* (email sentence completion); *cluster similarities* (online shopping recommender system); *discover patterns* (smartwatch activity trends) [see Figure 1 and supplementary materials]. Selection and curation of capabilities were not meant to be exhaustive; similar to prior work [84, 132, 138], our goal was to have a *good enough* subset to inspire ideation.

We conducted a 2-hour in-person workshop following the same structure as in the first workshop. However, this time we started by reviewing the AI capabilities and examples we had prepared in the form of slides during the introduction session (10 min). We used the slides as poster printouts to prompt ideation across each specific capability. For instance, talking about "email spam filter" as an example of binary classification (spam or not spam), we probed if we could envision use cases where classifying things as important or not important, or as urgent or not urgent could be useful. Ideation sessions were followed by impact-effort assessment and debriefing, as in the initial workshop.

4.1.3 Data Collection and Analysis. Workshops were audio and video recorded, and transcribed. The analysis included reviewing (1) the transcripts using interpretation sessions, and (2) workshop outcomes using affinity diagramming [66, 82], and the task expertise-model performance matrix [139] – a new assessment tool our team had created to assess the breadth of AI problem-solution space (see Figure 2b). This matrix broke down concepts into a two-by-two matrix based on two dimensions: **task expertise** (*how much human expertise or intelligence does this task require?*) and **model performance** (*what is the minimum quality needed for users to experience AI as useful?*). The analysis focused on identifying key themes for the concepts, challenges in collaboration, and the impact of design activities on workshop outputs. Two authors led the analysis, before sharing the results and insights with the entire study team for further review and discussion. We then iteratively discussed and restructured the emerging themes to seek agreement on interpretations across members.

4.2 Findings

In this section we present workshop results by describing (1) outcomes detailing the quality of concepts, and (2) our reflections on what worked, what did not work, and what was unexpected.

4.2.1 Workshop 1 Outcome. The first workshop was effective at getting all members of the team to ideate. However, the outcomes seemed to cover a narrow space. Our impact-effort assessment showed that the majority of our concepts were difficult to build, while only about half seemed relevant and useful for critical care medicine (Figure 2a). Our analysis of high-level brainstorming themes also indicated a lack of breadth: more than a third of concepts focused on clinical decision making, and another third described systems that automated documentation. A few of the concepts described new AI-enabled interactions. One concept described a system that forecasts expert disagreement. For example, it might predict that a nurse would not perform a specific assessment because they viewed the patient as not qualifying while the physician would want the assessment to have been performed. Another described an AI assistant that listens and transcribes conversations between clinicians.

Overall, our team collectively felt that the concepts were not very novel. Most of the concepts addressed existing interactions instead of proposing new ways of working. Concepts often described latent desires around trust, feedback, and explainability (e.g. *AI can take feedback on why it is wrong*); human-AI interaction forms (e.g. checklist, chatbot, recommendation system, conversational assistant); desired system behaviors (e.g. *recommendation is not intrusive, recommendation comes when ICU team is together*); and pain points (e.g. *placing orders is a burden; I want to eliminate and delegate tasks*).

Similar to the impact-effort assessment results, our task expertise-AI performance analysis showed that most of the concepts mapped to the upper right region (high expertise-excellent performance), missing the larger design space (Figure 2b). Concepts often required near-perfect AI performance or accuracy to be useful. For instance, anticipating clinician disagreement or predicting if a nurse will not perform an assessment can be useful *only* if the AI system can correctly capture 9 cases out of 10. The system would not be useful if it incorrectly flags situations or can only catch cases correctly once in a while. Our concepts also seemed too focused on situations with high uncertainty where the task is difficult even for highly trained experts (e.g., clinical decision making, anticipating potential disagreements).

Post-workshop reflection. Our brainstorming workshop was successful in that our multidisciplinary team generated many concepts for potential AI use cases. Data science and healthcare team members found the brainstorming exercise novel, as they had not previously engaged in formal, structured brainstorming or human-centered design perspectives. However, assessment of the workshop outcomes showed that the concepts were not of the quality we wanted. Our process was not generating any concepts that were easy to develop; *low hanging fruit* where moderate AI performance could generate value in the ICU. Some concepts did not require AI, and several called for data that does not exist. Reflecting on the outcomes, we set a new goal to move ideation towards *situations where moderate AI performance could still generate value.*

4.2.2 Workshop 2 Outcome. The second workshop led to concepts that mapped to a broader set of themes. This was one type of concept quality we were particularly focused on. Examples included AI systems that would improve coordination between clinicians (e.g. *generate a schedule for nurses and respiratory therapists for extubation*); systems that improved logistics and resource allocation (e.g. *predict which medications would be needed based on current patients and pre-order from pharmacy*); systems that inferred workload and effort, possibly in support of dynamic staffing (e.g. *classify patients as sick or busy*); systems that better support attention management (e.g. *classify alerts as urgent or not urgent*); systems that improve

Phase	Theme	Idea
W1	Decision support	Show outcomes from recent past patients
	Documentation	AI assistant that listens to clinician conversations
	Information retrieval	Summary of patient current state
	Patient-centric care	Insights on family care to enable ICU at home
	Personal informatics	Fitbit for clinicians: how am I doing?
	Team dynamics	AI recommendation system foresees areas of tension
	Workload management	Recommend how to better adjust workload
W2	Automation	AI suggests best billing code based on the patient note
	Coordination	Generate a schedule for nurses and respiratory therapists for extubation
	Decision support	Classify potential discharges based on vitals and most recent progress note
	Documentation	Recognize discrepancy in notes, i.e. doc A says X, doc B says not X
	Eligibility for EBP	Predict if patient is eligible for extubation
	Information retrieval	Learn what clinicians look at based on condition, prefetch to dashboard
	Patient-centric care	Predict when family would come, allow to prepare for family meeting
	Personal informatics	Listen to rounds, offer feedback on quality of leadership to team leader
	Reducing errors	Find orders in notes that are actually not ordered
	Resource planning	Predict what meds would be needed, pre-order from pharmacy
	Task acceleration	Predict and recommend orders typical for diagnosis
	Workload management	Classify patient as a busy patient or a sick patient

Table 2: High level themes and example concepts from first and second ideation workshops.

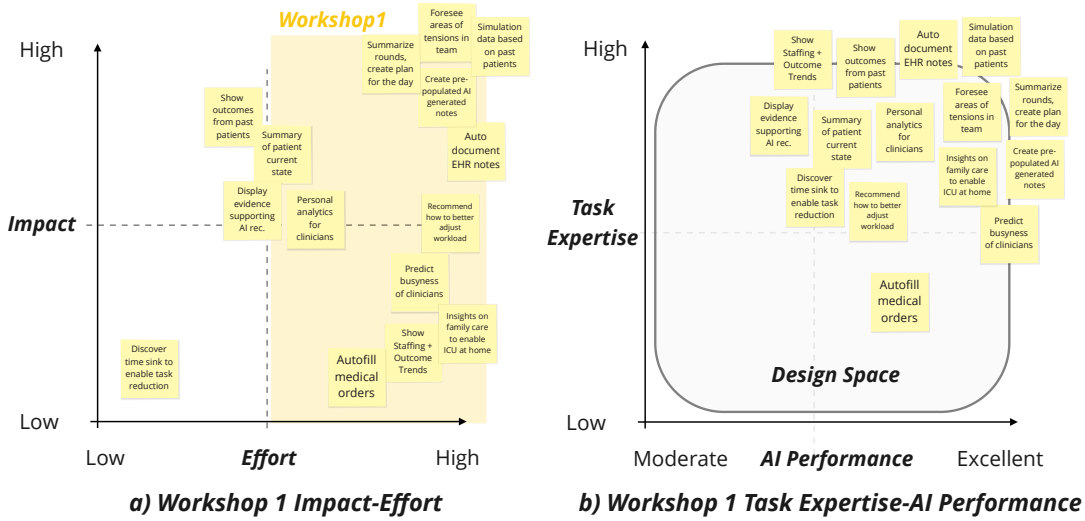

Figure 2: Our first workshop resulted in ideas that were technically difficult, some of which were clinically relevant.

efficiency, particularly around data entry and documentation (e.g. *predict and recommend orders typical for diagnosis*); systems that anticipate needed information (e.g. *learn relevant information based on patient conditions*).

In addition to these new themes, we generated concepts that built on the themes from the previous workshop, including decision support (e.g. *predict if the patient is eligible for extubation*); documentation (e.g. *generate a draft patient note based on available information*), and automation of menial tasks (e.g. *recommend best billing code based on the patient note*). Table 2 lists the high level themes and example concepts from each round of workshops.

Using AI capabilities and examples served as a springboard for our team to recognize situations where a capability could be useful to then effectively transfer that capability to a use case. For example, a nurse practitioner envisioned classifying patients into two groups, sick patients and busy patients. This mirrored the *classify things* capability. Sick patients typically require more attention. Busy patients included patients who needed many time-consuming procedures: *"Is this a busy patient? Or is this a sick patient? It would be useful for managing nursing tasks to tell the difference between a patient who's incredibly sick, but doesn't have a lot of tasks. ... [versus] if they have a lot of weeping wounds or something like that, that can make for a very busy patient." (Nurse 2, H8)* This concept hinted at the potential for more dynamic staffing or could be used to balance work difficulty and staff expertise across an ICU. Another capability, *observing and surfacing information*, spurred the concept

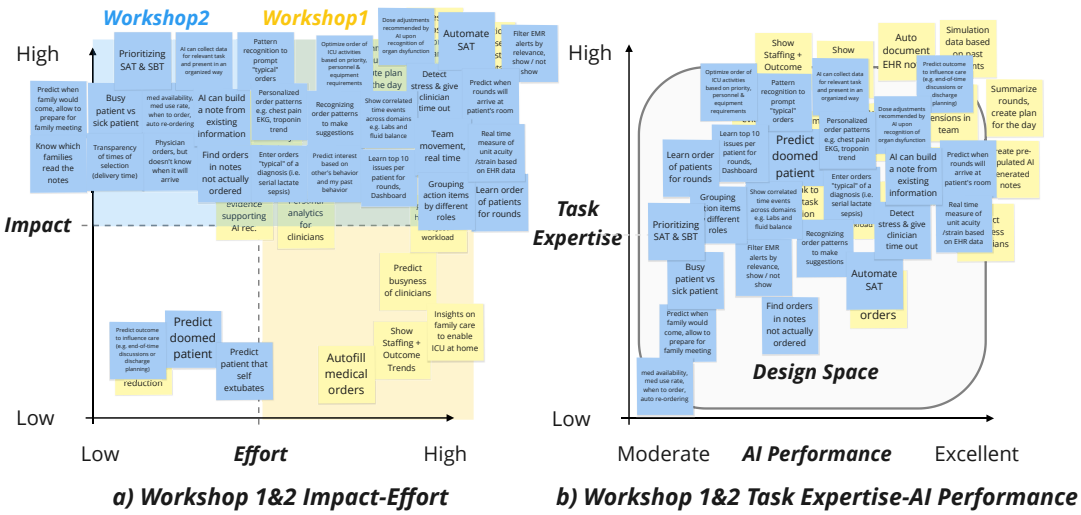

Figure 3: In the second workshop, concepts moved towards (a) low-effort and high-impact; (b) from high expertise-excellent performance to medium expertise-moderate performance.

of learning what EHR screens and information clinicians looked at based on patient condition in order to prefetch or highlight relevant patient history information on a dashboard.

In impact-effort assessment, our concepts moved towards the upper left quadrant: we were able to identify concepts that required low implementation effort with potentially high-impact (Figure 3a). The task expertise-model performance assessment also revealed that the concepts moved from high expertise-excellent performance to medium expertise-moderate performance (Figure 3b). For example, generating an ordered list of patients for rounds based on the uncertainty of what to do seemed relatively low-risk. A moderate quality, draft triage list is still better than no prioritization; the clinical team will still attend to all the patients in the ICU. Interestingly, in expanding the solution space towards situations where moderate AI performance could be useful, we moved beyond high-stakes situations with great uncertainty (e.g., clinical decision making) and produced concepts for relatively underexplored places (e.g., coordination, managing workload, anticipatory information retrieval).

4.2.3 Post-study Reflection. Discussing specific AI capabilities and examples prior to the workshop seemed to have a significant impact on the outcomes of ideation, yielding a broader design space where a mediocre, imperfect AI model would still provide enough value for clinicians. We also noticed that explicitly talking about AI capabilities provided our team with a shared language. Unlike the first round, most sticky notes described interaction concepts starting with capability verbs (e.g. *detect, recognize, classify, notice, predict, generate...*). Using this language, clinicians probed data scientists about technical possibilities. *"Can AI notice the sequence of orders? ... Can AI cluster tasks?"* Ideation became a collective conversation to discuss what would be doable, how that would produce value for users, and whether any relevant data was captured.

Although the quality of the concepts improved, we still encountered challenges. First, our assessment showed that while our concepts were grounded in what's technically possible, only a few of them were implementable using our specific ICU dataset. Most concepts required additional data collection or instrumentation (e.g. tracking clinician clicks in UI to learn and pre-fetch information to dashboards). In some cases, the data existed but it was not in our dataset (e.g. unstructured text from clinical notes), rendering our concepts infeasible unless we sought out more and different data. Overall, the ideation exercise was valuable for informing data collection for future implementations, but we were ignoring the value of our ICU dataset in our ideation. We needed concepts we could build using our data to create immediate value for clinicians.

We also noticed that similar to other healthcare innovation research [133], we had a tendency to attribute familiar interaction forms, such as alerts, to specific capabilities and concepts based on past experiences. For instance, while we liked the concept of classifying patients, we always seemed to imagine this as an alert or a reminder. Given the well-known research on alert fatigue and clinician burnout [21], this seemed problematic. Our fixation on existing forms bound to a capability posed a threat to ideation, as the team would dismiss concepts based on known problems with the familiar forms. As prior research reported [132], we found ourselves trying to separate the inference (e.g., predicting that a patient would need a scan) from the interaction (e.g., recommending the action to a clinician or proactively ordering a scan).

Relatedly, rapid ideation resulted in surface level concepts that require further exploration. For instance, clinicians liked the concept of having a ranked list of patients to visit during rounding. However, the criteria needed to prioritize patients was not clearly defined: should it be based on sickness level (see sickest patients first) or patient uncertainty (patients where it was least obvious what to do)? In order to more effectively assess the concepts and select candidates for development, we needed more detail on what the concept was and how it might work in terms of data requirements and the form of the AI output clinicians would encounter.

Finally, following the second workshop, discussions on how to move forward surfaced confusions and a need for increased communication within the team. While the HCI team perceived the second workshop as a success –especially from a methodological point of view– the shift in the quality of ideation was not obvious to the rest of the team. Conversely, the data science and healthcare team members found the exercise to be repetitive. The clinical team lead expressed confusion over the activities in the second workshop, probing the reasoning behind generating concepts from scratch instead of building on the existing ideas from the first workshop. To resolve concerns, the lead HCI researcher presented the post-workshop assessment of concepts, clarifying how the quality and breadth of ideation has shifted. The team then reached a consensus that the next best step would be to select a subset of ideas that could be grounded within our ICU dataset for further detailing and assessment.

5 PHASE 2: PROBLEM FORMULATION

As we moved from ideation to problem formulation, we set three goals. First, we wanted to leverage the unique properties of our dataset, and ground our concepts in what we could realistically build. Second, we wanted to separate interaction form and AI inference when discussing concepts. Third, we wanted to deeply explore some of the concepts to have more mature conversations on their feasibility, desirability, and potential implications.

5.1 Method

We chose to conduct an additional design workshop that focused on problem formulation. Similar to the phase 1 study, we conducted a 2-hour in-person workshop for detailing a subset of 12 concepts. Below, we first describe how we prioritized and selected the subset of concepts prior to the workshop. We then detail the artifacts prepared for the workshop and the set of activities.

5.1.1 Concept Prioritization. We had three criteria when selecting concepts for further development. First, we prioritized concepts based on data availability, choosing concepts that could be built using our ICU dataset. Second, we sought to cover a breadth of the design space, selecting concepts where moderate-to-good performance AI could produce medium-to-high value. Finally, we included concepts that matched our team's research interests and expertise, excluding some concepts in subspecialty AI areas (e.g., natural language processing or computer vision-based concepts). The selected concepts included: anticipatory pre-ordering of medications; predicting medication time-to-delivery; generating a prioritized list of nurse assignments; identifying sick or busy patients; forecasting unit acuity; generating an ordered list of patients to see for rounds; predicting the eligibility of patients for extubation from mechanical ventilators; generating a coordinated schedule for extubation; identifying clinician workload patterns; identifying bias in clinical orders; predicting typical orders for diagnoses; and discovering the sequence of tasks.

5.1.2 Workshop Preparation. Prior to the workshop, the lead HCI researcher worked on numerous representations to untangle the inference produced by an AI model, the data needed to build the model, and the form of the AI output clinicians would encounter. Over several discussions, the team critiqued and iterated on the alternative artifacts. After rounds of iterations, we arrived at a new abstract representation: *the Do-Reason-Know worksheet* (Figure 4). Each section respectively captures the interaction (do), model reasoning and inference (reason), and data requirements (know).

The worksheet builds on the classical input-model-output representation commonly used in machine learning [47], yet it furthers the existing artifacts in two aspects. First, it captures both the inference and the delivery of the inference for separating the model behavior (e.g. rank patients) from the interaction behavior (e.g. present a list where critical patients are displayed at the top). Second, it balances the model-centric view with a user-centric view by flipping the starting point (end user interaction instead of AI input or output), and embedding the desired system behavior into problem formulation from the beginning. In preparation for the workshop, we pre-populated the worksheets with the concept names and any other relevant information that was discussed in prior workshops (e.g. a potential data source our team had referred to related to a particular concept).

5.1.3 Workshop Activities. We conducted a 2-hour in person workshop. The workshop kicked off with a short review of the worksheet and the 12 concepts we pre-selected (15 min). Then, we divided into two groups, where each group collectively discussed and detailed 6 concepts (90 min). We used worksheet printouts as a starting point and detailed each section by adding sticky notes. For instance, when deliberating on *predicting whether a patient might need a certain procedure (e.g. surgery, intubation)*, we discussed if the time of a procedure is documented and whether there were relevant actions or events we could use as proxies (e.g. bleeding prior to surgery). We concluded with a brief reflection and discussion on the next steps (10 min).

5.1.4 Data Collection and Analysis. We audio and video recorded and transcribed the workshop. We documented the worksheet printouts, and analyzed the transcripts and artifacts using the same methods as in Phase 1 (see section 4.1.3).

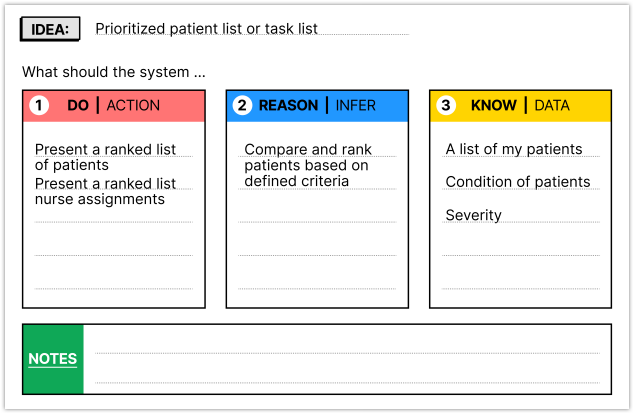

Figure 4: Do-Reason-Know worksheet enabled us to detail each idea in terms of action, model reasoning, and data.

5.2 Findings

We first present insights from the workshop capturing our process of problem formulation. We then reflect on the use of the Do-Reason-Know worksheet in concept detailing.

5.2.1 Workshop Outcome. One of our goals was to focus on low-risk, medium-value concepts. Throughout the workshop, we reworked our concepts in a way that reduced the required model performance to help us identify relatively simple, low-risk AI concepts. We repeatedly asked *"Is there a simpler, dumber version of this concept that is still 'good enough' to produce value?"* Below, we share three concepts to illustrate how this approach helped us effectively formulate concepts.

Predicting if a mechanically ventilated patient is eligible to receive a breathing trial, instead of predicting if the patient should be extubated. Liberation from mechanical ventilation is a complex process that requires coordinated actions of nurses, respiratory therapists, and physicians. It involves two integrated actions. Typically, the nurse assigned to a specific patient will perform a *Spontaneous Awakening Trial (SAT);* they will cut off a patient's sedation and observe if they can tolerate being awake. Next, the respiratory therapist, who is typically in charge of making changes to the ventilator settings, will perform a *Spontaneous Breathing Trial (SBT).* They will suspend breathing support and observe how well patients take over their own breathing. These assessments allow the team to decide if a patient can be extubated (liberated from a ventilator).

Remaining on a ventilator is associated with several adverse outcomes including delirium, pneumonia, and heart damage; however, extubating the patient and taking them off the ventilator too soon leads to another host of problems [52, 64, 83]. When one of the steps gets missed (SAT and SBT), then the clinical team lacks the information to make a decision about extubation, meaning the patient remains on the ventilator for another day.

Our initial concept around patient extubation focused on *predicting if a patient will successfully extubate* and *discovering the right amount of sedation for a patient on a ventilator.* These are hard problems that need excellent model performance and very high quality healthcare data, which may not exist. During our discussions, clinician team members shared that physicians can become risk averse when extubations fail. They speculated that this might result in patients remaining on a ventilator longer than needed.

With this in mind, we turned our attention to the execution of SAT/SBT procedures instead of the clinical decision making for patient extubation. This led the concept towards *predicting a patient's eligibility to receive SAT/SBT.* This is a comparatively low-risk, moderate-performance, and medium-value concept, as it focuses on an intermediate, safe-to-perform action rather than a critical decision.

Predicting medication availability and anticipatory ordering. One of the promising concepts that emerged from our ideation was predicting what medications would be needed based on the patient conditions in the unit. The concept was inspired by Amazon's anticipatory shipping [110] –an AI capability and example that came up during capability-based ideation workshop– where the AI system would keep track of the inventory and pre-order medications to reduce time and uncertainty.

During problem formulation, clinician team members shared that this would be really useful for custom mixed antibiotics: *"Sometimes you say 'Antibiotics. Now!' and two hours later it still hasn't arrived." (Physician 1, H1)* They noted that delays are more likely to happen in busier wards, which can be deadly [45]. However, clinicians were also cautious as the incorrect predictions might lead to unused medications, and therefore waste.

We broke down this concept into several lower risk concepts. First, instead of preordering, the predictions could be used only to inform the pharmacists so that they have a sense of what to expect. Second, we could instead predict time-to-medication to provide feedforward to the clinical team when placing orders. Third, a simpler approach could check for antibiotic dosing errors to prevent delays:

> ***Physician 2:*** *"I want this antibiotic for my patient. When the pharmacist finally gets to it, they say, you ordered the wrong dose. Because this patient is this size, this weight and has this renal function. Something smart would be able to figure that out, like smart dosing." (H2)*
> ***Data Scientist 1:*** *"That's a lot easier to do. We have that history of conditions, and what was given to [patients], so maybe these kinds of predictions." (DS1)*

5.2.2 Use of the Do-Reason-Know Worksheet. The worksheet helped to scaffold conversations around data dependency, model behavior, and interaction behavior. It allowed us to express concepts in a more refined way as we moved from sticky note concepts to more fleshed out problem formulations. It prompted us to further probe each concept in terms of how it would generate value for clinicians, and which features in our dataset could drive it, if at all possible. For instance, when discussing what *patient priority* means:

> ***Physician 4:*** *It's a two by two table. There are sick people that if you do the things you need to do, they're going to be just fine. And then there's the sick people who are uncertain. I need to pay attention to this patient in the next four hours because if I don't, six hours from now, they might be dead. ... [It would be great if] it was clear who those patients were, and you didn't have to take 15 minutes to figure that out. (H4)*
> ***HCI Researcher 1:*** *What information helps you determine which category that patient falls into? (HCD5)*
> ***Physician 4:*** *I look at what drips they're on, what's their vent settings. You'd be looking at the amount of drip titration, certain kinds of orders, certain kinds of labs, maybe some radiology findings. I think you can observe some of that in the data. (H4)*
> ***HCI Researcher 1:*** *How accurate do you feel like your rankings are after you spend fifteen minutes? (HCD1)*
> ***Physician 2:*** *There can be surprises, but I'm relying on my team to give me a better idea. (H4)*
> ***HCI Researcher 4:*** *Do you think it would be useful? At which point this would be most useful? (HCD4)*
> ***Physician 2:*** *The idea is to reduce the cognitive load on the physician. That's probably most useful at the beginning of the day, maybe at the end of the day when we switch shifts, handing off to the other person. If there*

> *was a tool there, I might check it once or twice throughout the day like, has anything changed? (H2)*
> ***Data Scientist 1:*** *Presumably in the algorithm, we could do it every four hours. (DS1)*

Describing the concept with this level of detail made it clear this would function as two separate two-class classifiers. Each patient would be classified as *not-sick* or *sick*, and they would be classified as *certain of what to do* or *uncertain of what to do*. Interestingly, as the model capability and reasoning became clear, our discussions moved towards:

(1) **Model performance:** How accurate or robust do the predictions need to be?
(2) **Point of interaction:** When, where, and how the inference should be delivered to produce value? (e.g. *are predictions available 15 minutes before or the night before?*)
(3) **Risk:** What are the consequences of errors? (i.e. *false positives and false negatives*)
(4) **Data quality:** Is the training data trustworthy? Is it likely to introduce bias?

Specifically, the worksheet helped with the three challenges we previously encountered. First, it allowed us to collectively define and formulate AI experiences in a way that is grounded in our dataset. Second, it allowed us to free up our concepts from existing forms by separating the interaction, AI capability, and data. Third, it informed our design deliberation and supported a deeper discussion of the concepts before starting model building and prototyping. For example, when discussing the concept *predicting typical orders for diagnoses*, one physician likened this to a personalized contacts list in email clients, where typing upon a contact name would present the most frequent contact at the top. The personalization aspect raised some concerns: would the medication orders be based on an individual clinician's previous orders or based on a group of clinicians' orders? Physicians seemed to prefer a personalized system, which seemed more complex and costly (both in terms of model building and continuous learning). These deliberations helped us weigh cost-value tradeoffs throughout problem formulation.

Our third workshop had an additional, unexpected benefit: our discussions helped our team to reveal existing or potential problems in our dataset. For instance, one of our ideas was around predicting patient eligibility for extubation from a mechanical ventilator to help clinicians plan for extubation. While exploring potential features in our data, we discussed whether we could use Riker scores, a numeric score for documenting the level of a patient's sedation level and consciousness. When discussing this concept, healthcare members shared that Riker score data were not trustworthy. The scores nurses entered into the EHR did not always reflect the actual level of sedation. This problematic data did not impact the quality of care as clinicians looked at the patient before making a decision. They did not make sedation decisions based on what was captured in the EHR. Thus, they never fixed this data entry problem. Interestingly, this issue is neither reported nor speculated in medical literature. Uncovering this insight early on in the process helped our team avoid using data features that clinicians did not trust.

5.2.3 Post-study Reflection. The problem formulation workshop with the focused worksheet activity helped us detail our concepts for further development. Following this workshop, we decided to sketch out some concepts in detail to elicit initial user feedback. Notably, the workshop debrief revealed many insights into the felt experience of our team. For instance, the clinical team lead found the workshop series valuable from a portfolio building and de-risking point of view: *'In [healthcare ML research] there is a lot of inertia towards low-risk, low-reward areas that doesn't move the needle in a meaningful way. This exercise is really valuable because people can replicate these methods to identify lower-risk yet high-reward ideas that are worth doing. Every research portfolio should have a mix of those.' (H1)* Reflecting on how the exercise can be improved, some clinicians shared that involving a broader set of stakeholders would be more helpful: *'It might be useful to have in the room like somebody from hospital management, somebody from pharmacy ... to help fill in some of the gaps, [as we have] been making some assumptions.' (H2)* Finally, all data science team members expressed that they found the third workshop the most beneficial. It seemed to help them to gain a deeper understanding of clinical domain knowledge in relation with the data: *"It's great to hear how and where the data is coming from." (Data Scientist, DS2*). After the workshop, several data science team members shared additional concepts or ideas on implementation details with the team based on the insights our discussions sparked.

From a methodological perspective, using a combination of impact-effort matrix and task expertise-AI performance matrix, along with the Do-Reason-Know worksheet allowed us to quickly sort out ideas that our team was most interested in. However, in hindsight, we noticed that dimensions, such as impact and effort, can be even more granular for a more rigorous concept assessment. For example, questions around effort included *'is there any data available?'*, *'how much work is needed for data cleaning or anonymization?'*, and *'how easily can we measure and validate AI outputs?'* Moreover, the AI performance and effort (feasibility) seemed related; we repeatedly asked *'what level of performance is needed?'* and based on that *'how difficult is it to achieve that performance?'*. We also noted two other critical dimensions that we have not delved into: financial viability *('how expensive is this model to build and run?'*, *'how much return on investment (ROI) is it likely to generate')* and potential responsible AI issues *('are there issues around privacy, fairness, data bias?')*. We reflected that capturing these dimensions in a more nuanced manner can inform the future iterations of the Do-Know-Reason worksheet (e.g., similar to datasheets [43, 105], a comprehensive 'AI concept template' for concept proposals).

6 PHASE 3: SKETCHING AND CO-DESIGN

Following ideation and problem formulation, we chose to further develop the concept of *predicting if a mechanically ventilated patient is eligible to receive the SAT/SBT protocol*. We engaged in a concurrent model development and interaction design process. The clinician and data science team members carried out the data and model work, and the HCI team members conducted co-design sessions with end users. In this section, we provide a brief overview of our sketching and co-design process to illustrate how we moved from ideation towards sketching and concept refinement, envisioning how clinicians might interact with an AI system.

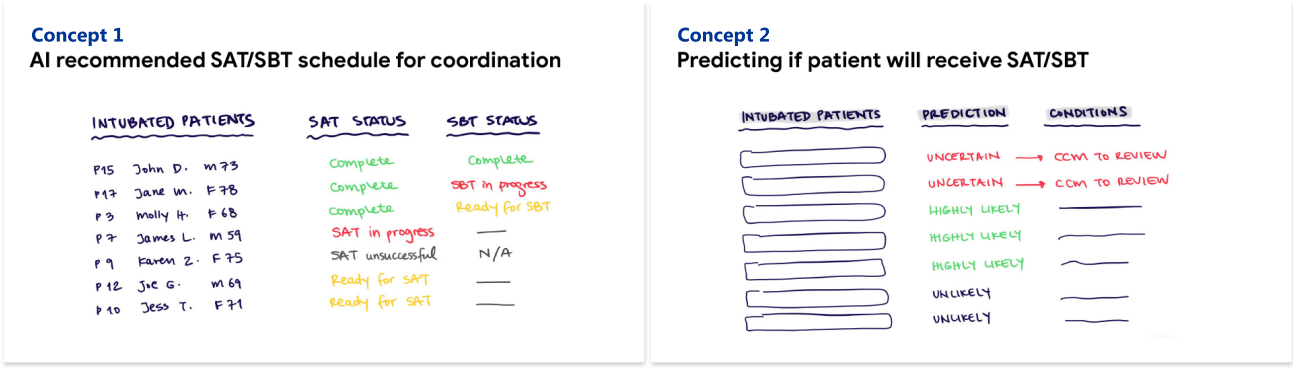

Figure 5: Two low fidelity sketches detailing the concept of *predicting if a mechanically ventilated patient is eligible to receive the SAT/SBT protocol.* We used the sketches to conduct co-design workshops to elicit feedback from nurses and respiratory therapists.

6.1 Method

6.1.1 Concept Sketching. We created two low fidelity concept sketches detailing a shared dashboard for nurses and respiratory therapists (RTs) to support them in executing the SAT/SBT protocol for mechanically ventilated patients. The first concept displayed a dashboard with an AI-generated SAT/SBT patient schedule for better coordination (Figure 5a). The second concept displayed a dashboard that predicted if a patient will receive an SAT/SBT based on past data, and ranked the patients based on uncertainty. In this concept, high uncertainty patients were displayed on top, so that the care team could resolve uncertainties at the beginning of the morning shift (Figure 5b).

6.1.2 Co-Design Workshops. We conducted four co-design workshops with 6 RTs and 5 nurses. In each session, we had at least one nurse and one RT participant. We recruited participants through a mix of purposive and snowball sampling [51], first reaching out to our contacts at collaborating hospitals, then expanding this set by asking participants to share relevant contacts. Workshops were conducted in-person, and facilitated by the lead HCI researcher. We first probed participants about their current practices for executing the SAT/SBT protocol. We then shared the concepts as print outs, asking them to reflect whether and how these could be useful. We provided markers and pens for participants to directly edit and comment on the concepts. Each session lasted approximately 2 hours. Participants were compensated $250 for their time. The study was approved by our Institutional Review Board.

6.1.3 Data Collection and Analysis. All workshops were audio and video recorded, and transcribed verbatim. We also documented the printouts that recorded participants' notes. We analyzed the data using the same methods described in Phase 1.

6.2 Findings

6.2.1 Workshop Outcome. Overall, our participants perceived the concepts as valuable. They reflected that having a shared dashboard that pre-assessed a patient's eligibility for the protocol and documented any contraindications would help them plan for patients. Several participants desired the system to not only show the patient eligibility, but also the longitudinal SAT/SBT history: *'Specifically, did they meet the criteria? How long were they on it? What was the contraindication? Was this SBT done? What were the settings? Was it successful? If not why?'* They also indicated that the system could offer meaningful category labels to indicate why a patient was categorized as ineligible: *'A good category [for ineligible patients] would be seeing condition A, if they were called for an emergent reason [such as] airway protection, drug overdose.' (RT1)*

Both nurses and RTs reflected that patients who have high uncertainty are often deprioritized as the uncertainties tend to go unresolved, resulting in eligible patients not receiving the protocol. An RT reflected that flagging these patients would be useful for the care team to review: *'If the nurse charts that their neuro function is not normal, it's probably uncertain to me, the doctor needs to review. So if those were put in the algorithms and sorted out, I can tell who I'm going to see first.' (RT5).* However, some participants indicated that they would not trust an algorithm-based patient prioritization. They expressed a desire for the involvement of the physician, who could review this draft list to adjust the patient priority based on their goals. Finally, participants expressed that knowing high risk patients –patients who are most likely to fail the breathing trial– might be useful for planning and coordination: *'If you know every one of your patients [who] is going to be absolutely terrible when you SBT them, you might want to do all your other SBTs first, and then get them last to make sure your nurse is with you in the room.' (RT3)*

6.2.2 Post-study Reflection. Initial feedback we gained from nurses and RTs informed both the interaction design and modeling work for this concept. Moving forward, we aim to iterate on the concept to convey both patient trajectory and priority –places where ICU clinicians think AI can help [37]– to help clinicians consider and perform evidence-based protocols.

7 DISCUSSION

Our work have explored facilitating early stage AI ideation and problem formulation – an opportune moment for involving domain stakeholders in identifying the right thing, or a good enough thing, to build [95]. We built on the prior observation that effective

innovation teams brainstorm *many* AI concepts by using AI capabilities and examples, before selecting a concept to further develop [134, 138, 140]. We share a case study detailing how our multidisciplinary team effectively engaged in brainstorming AI concepts for the ICU. Below, we reflect on how this approach can generalize to high-stakes, critical domains to reduce the risk of developing unwanted technology. We detail what challenges remain for moving from ideation to prototyping, and discuss open research questions and the limitations of this work.

7.1 Towards Participatory AI in High-Stakes, Critical Domains

Researchers have called for participatory approaches to AI for engaging a broad set of stakeholders in early phase brainstorming to explore AI's potential value and risks in high-stakes, critical domains [20, 27, 50, 68]. However, it remains unclear how, when and to what extent this would be possible [9, 14, 28, 104]. We took a step towards this direction in the context of healthcare. This is a relatively challenging design space to navigate, as we did not bind our ideation to specific AI mechanisms (e.g., clinical NLP) or interaction forms (e.g., AI-assisted diagnosis). We approached this challenge by holding design workshops, hoping that by bringing data science, HCI, and domain experts together, we could elicit what is clinically relevant and feasible. However, simply asking clinicians what would be most valuable did not prove effective: concepts were largely unbuildable or unwanted. We suspect that following a user-centered process has unintentionally led our team to focus on problems that do not need AI – points of great uncertainty or edge cases where AI is not likely to work. Additionally, traditional rules of brainstorming, such as letting go of technical limitations, seemed to exacerbate the problem of generating unbuildable concepts.

In search of a more effective process, we took a step towards matchmaking [10]. Starting with AI capabilities and examples, and then asking clinicians if they recognize situations where capabilities would be useful and where moderate performance could create value, led to more effective ideation. It resulted in a broader coverage of the problem-solution space, leading to technically achievable and clinically relevant concepts. Capability abstractions and examples scaffolded clinicians' understanding of what AI can do, and gave our team a shared language to discuss what would be possible. In addition to discovering value, engaging domain experts in concept generation and assessment helped us surface potential risks. We were able to identify which data features we should *not* use, data that could not be trusted.

This provides a glimpse into what effective ideation and problem formulation might look like, and how it might help situate AI in high-stakes, critical work contexts. Future research should investigate whether this approach might generalize within and beyond healthcare. Does reviewing AI capabilities and examples with moderate performance help domain experts systematically yield high-impact, low-risk concepts? How does the selection of examples and capabilities impact the quality of generated concepts? Comparing the two brainstorming approaches – workshop 1 and workshop 2 – poses additional challenges: it is difficult to assess whether there is an interaction or order effect, since starting with AI capabilities will immediately sensitize the team to what AI can do. We encourage HCI and design researchers to share first-person accounts and case studies swapping and modifying these approaches to ideation to guide our community in constructing a better design process as well as new educational exercises. Recent work (e.g., [86, 87]) provides great starting places for this line of inquiry.

Our work focused on clinicians as the domain stakeholders, yet there are many critical stakeholders in healthcare including patients, caregivers, hospital managers, insurance companies, and regulatory bodies. How could we blend matchmaking with participatory design where all stakeholders can meaningfully engage? What is the earliest point in the design and development process to engage domain stakeholders? While we started our project post-collection, we suspect that generating AI concepts *prior* to data collection could inform the collection of high quality data in the first place. Recent literature suggested proactive and intentional data collection practices through pre-collection planning and documentation [56, 90, 97, 118, 146]. Future research can build on this line of work by engaging diverse stakeholders in *designing data* to inform what should and should not be collected.

7.2 Moving from Ideation to Prototyping

Sketching –generating many different ideas in order to discover the right thing to make– and *prototyping* –making the thing at increasing levels of fidelity to refine it into being– are cornerstones of HCI practice [17]. Envisioning and prototyping AI experiences pose many unique challenges for innovation teams, especially at the early stages of ideation, problem formulation, and project selection [140]. Throughout our ideation process, we utilized several resources and artifacts that can serve as a launching pad for cross-disciplinary, collective ideation. To summarize, we used:

- *A set of AI capability abstractions and examples, detailing what AI can do and how it has previously produced value, especially with moderate performance.* These capabilities offered a starting point for discussing whether AI could solve a problem that particularly seemed like a good match. The capability abstractions provided a shared language and encouraged our team to bring up more examples throughout the ideation.
- *A combination of assessment matrices delineating task expertise-model performance and impact-effort.* Noticing the interplay between these dimensions helped us map the design space, and guided our search and prioritization.
- *A worksheet capturing the interaction, model reasoning, and data.* The Do-Reason-Know worksheet enabled us to effectively enrich an concept and understand its potential impact and limitations. It helped us to separate interaction form and model behavior. It also supported a deeper discussion on the data source, allowing us to flag data features that were unavailable, unreliable, or potentially biased.

Starting our project, one of our goals was to identify *low hanging fruit* – situations where simple AI interventions could improve clinical work. Based on prior research highlighting the value 'imperfect AI' can bring [12, 73] as well as our own work, we focused on *AI model performance* to sensitize our team to situations where moderate model performance can still bring enough value. Additionally, we repeatedly probed team members to think of simpler versions of concepts. This explicit consideration opened up a design

space beyond the automation of mundane tasks or augmentation of critical tasks. It surfaced things that humans would never do as it would not be worth their time for the return value *(e.g. predicting patients with high uncertainty to receive a clinical protocol, predicting what medications will be needed for patients to reduce pharmacy wait times)*. These low-risk situations present a great entry point for introducing AI in healthcare, which can inform our understanding of how people can and should collaborate with AI before deploying AI in high-stakes situations, such as decision making.

While these resources scaffolded and improved our ideation process, challenges remain in selecting a concept for further development. How do we analyze, compare, and select concepts in a more systematic way? Can we engage a broad set of impacted stakeholders, including patients and other clinical roles, to anticipate risks, fairness issues, and potential harm? What are some critical dimensions that are not captured by current assessment tools? Recent research uncovered assessment matrices industry practitioners created to assess and prioritize AI-enabled product features, which captured risk, frequency of use, and accuracy [140]. Similarly, our discussions surfaced risk of errors, data quality, acceptable model performance, and timing and presentation of information as key aspects to consider. Future research should investigate developing new assessment tools that move beyond typical metrics (e.g. feasibility, desirability) to capture the complexity of AI concept proposals. Moving from ideation to parallel prototyping –both experience prototyping and prototyping with data– our community would benefit greatly from having a robust assessment and selection process.

7.3 Open Research Questions

Our study revealed several open research questions. Below, we detail two challenges that merit further study.

7.3.1 How much AI knowledge is needed for domain experts to engage in ideation? Recent HCI research has explored the critical role domain experts play in AI development processes, especially in high-stakes domains [22, 106]. Researchers note that AI developers cannot readily elicit input from domain experts, and are often compelled to hold AI education sessions to span communication gaps [74, 88, 96]. What kind of AI literacy is needed for domain experts to effectively participate in AI envisionment? What kind of AI resources can help domain experts in engaging in ideation? How can we extend the set of AI capabilities and examples for use in other domains and contexts? Developing and assessing resources for stakeholder engagement in ideation, problem selection and formulation marks a clear direction for future research.

7.3.2 What makes an AI example "good"? Our research surfaced a key question: what makes an example 'good'? How do we select *a good enough subset of examples* that illustrate a breadth of AI capabilities and value propositions? We approached the capabilities and examples only as a subset to sensitize our team to think of other examples and capabilities. We also paid attention to the level of AI performance in each concept, and made sure to include examples where moderate performance created value. Interestingly, our team responded well to this approach and started drawing from other examples based on each member's prior experience. This approach of having *"a good enough subset"* was effective, as it would be incredibly challenging to try to represent and go through all AI capabilities. In this work, we utilized the AI Brainstorming Kit [139] to select capabilities and examples. Future research should investigate the use of this resource and others (e.g., [60]) to explore how selecting a subset of capabilities impact ideation, and how teams can effectively curate and review capabilities and examples.

7.3.3 Can early phase ideation and assessment address the high AI failure? User-centered design and participatory design grew out of HCI research addressing high rates of software product failures in the 1970s and 80s. Software engineers would select applications and start writing code; the idea of investigating what users want, need, and fear *before* making software was non-obvious. We see parallels between early software development and current AI product development. Recent research echoes this: industry product teams report repeatedly experiencing AI project failures due to working on the wrong problem [140]. We suspect that HCI experts can play a key role in AI development by helping teams find *the right AI thing to build while reducing the risk of potential harm.* This is especially true in high-stakes contexts, such as healthcare and public sector [20, 22, 27, 68], where AI teams do not seem to ideate on their own. HCI routinely facilitates the process of technology innovation between multiple stakeholders to reduce the risk of developing products and services nobody wants. What is uniquely difficult about facilitating AI ideation? We strongly encourage researchers to explore the role of HCI in facilitating collective AI ideation and problem formulation.

8 LIMITATIONS

Our study had two limitations. First, we focused solely on sketching. While we are in the process of prototyping and model building for a few of the selected concepts, we do not claim that all concepts we generated are feasible, valuable, or novel in practice. Instead, we assess the perceived difficulty and perceived value of the concepts. This trade-off between sketching and prototyping was intentional, as our focus was on broadly exploring many concepts. Future research should investigate ideation followed by parallel prototyping of multiple concepts to assess the impact and technical effort required for implementation. Second, we do not know if there was an order effect on our ideation process. Future work should conduct controlled studies to compare the user-centered *and* tech-centered approach we propose with traditional, user-centered brainstorming.

9 CONCLUSION

This paper presented a case study of early phase AI innovation capturing multidisciplinary concept ideation and problem formulation in the context of healthcare. Our work offers insights into how teams might structure their design process to effectively explore AI's problem-solution space and engage domain experts in ideation. We documented our case with high-fidelity, detailing the challenges we encountered and our emergent solutions. Our case suggests that starting ideation with AI capabilities leads to broader exploration of the solution space, and sensitizing teams to the level of AI performance needed surfaces lower risk concepts where moderate AI performance can still be useful. It also suggests that detailing concepts in terms of data, inference, and form helps to rapidly identify problems and makes concepts more pliable to interrogate easier,

simpler versions. While we conducted this work in the context of intensive care, we suspect this ideation and problem formulation process would generalize to many AI innovation projects that involve domain experts. Through this work, we hope to deepen the discussion on HCI's role in engaging multidisciplinary teams and stakeholders in AI ideation.

ACKNOWLEDGMENTS

We thank the participants in this work for their time and valuable input on the dashboard concepts. This material is based upon work supported by the National Science Foundation under Grant No. (2007501) and work supported by the National Institutes of Health (R35HL144804). The first author was also supported by the Center for Machine Learning and Health (CMLH) Translational Fellowships in Digital Health. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or the National Institutes of Health.

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–18.

[2] 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*. 1–13.

[3] Apple. 2019. Human Interface Guidelines: Machine Learning. https://developer.apple.com/design/human-interface-guidelines/technologies/machine-learning/introduction/

[4] Amid Ayobi, Jacob Hughes, Christopher J Duckworth, Jakub J Dylag, Sam James, Paul Marshall, Matthew Guy, Anitha Kumaran, Adriane Chapman, Michael Boniface, et al. 2023. Computational Notebooks as Co-Design Tools: Engaging Young Adults Living with Diabetes, Family Carers, and Clinicians with Machine Learning Models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.

[5] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M Vardoulakis. 2020. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–12.

[6] Andrew Bell, Oded Nov, and Julia Stoyanovich. 2023. Think about the stakeholders first! Toward an algorithmic transparency playbook for regulatory compliance. *Data & Policy* 5 (2023), e12.

[7] 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*. 1–14.

[8] Andrew BL Berry, Catherine Y Lim, Tad Hirsch, Andrea L Hartzler, Linda M Kiel, Zoë A Bermet, and James D Ralston. 2019. Supporting communication about values between people with multiple chronic conditions and their providers. In *proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[9] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Díaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the People? Opportunities and Challenges for Participatory AI. *Equity and Access in Algorithms, Mechanisms, and Optimization* (2022), 1–8.

[10] Sara Bly and Elizabeth F Churchill. 1999. Design through matchmaking: technology in search of users. *interactions* 6, 2 (1999), 23–31.

[11] Sander Bogers, Joep Frens, Janne Van Kollenburg, Eva Deckers, and Caroline Hummels. 2016. Connected baby bottle: A design case study towards a framework for data-enabled design. In *Proceedings of the 2016 ACM conference on designing interactive systems*. 301–311.

[12] Claus Bossen and Kathleen H Pine. 2023. Batman and Robin in Healthcare Knowledge Work: Human-AI Collaboration by Clinical Documentation Integrity Specialists. *ACM Transactions on Computer-Human Interaction* 30, 2 (2023), 1–29.

[13] Margarita Boyarskaya, Alexandra Olteanu, and Kate Crawford. 2020. Overcoming failures of imagination in AI infused system development and deployment. *arXiv preprint arXiv:2011.13416* (2020).

[14] Tone Bratteteig and Guri Verne. 2018. Does AI make PD obsolete? exploring challenges from artificial intelligence to participatory design. In *Proceedings of the 15th Participatory Design Conference: Short Papers, Situated Actions, Workshops and Tutorial-Volume 2*. 1–5.

[15] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.

[16] Eleanor R Burgess, Ivana Jankovic, Melissa Austin, Nancy Cai, Adela Kapuścińska, Suzanne Currie, J Marc Overhage, Erika S Poole, and Jofish Kaye. 2023. Healthcare AI Treatment Decision Support: Design Principles to Enhance Clinician Adoption and Trust. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.

[17] Bill Buxton. 2010. *Sketching user experiences: getting the design right and the right design*. Morgan kaufmann.

[18] Miguel Cabral Guerra, Deedee Kommers, Saskia Bakker, Pengcheng An, Carola van Pul, and Peter Andriessen. 2019. Beepless: Using Peripheral Interaction in an Intensive Care Setting. In *Proceedings of the 2019 on Designing Interactive Systems Conference*. 607–620.

[19] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. " Hello AI": uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. *Proceedings of the ACM on Human-computer Interaction* 3, CSCW (2019), 1–24.

[20] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2021. Onboarding Materials as Cross-functional Boundary Objects for Developing AI Assistants. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–7.

[21] Jared J Cash. 2009. Alert fatigue. *American Journal of Health-System Pharmacy* 66, 23 (2009), 2098–2101.

[22] Hao-Fei Cheng, Logan Stapleton, Ruiqi Wang, Paige Bullock, Alexandra Chouldechova, Zhiwei Steven Steven Wu, and Haiyi Zhu. 2021. Soliciting stakeholders' fairness notions in child maltreatment predictive systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–17.

[23] Vanessa Cobus, Bastian Ehrhardt, Susanne Boll, and Wilko Heuten. 2018. Vibrotactile alarm display for critical care. In *Proceedings of the 7th ACM International Symposium on Pervasive Displays*. 1–7.

[24] Ned Cooper, Tiffanie Horne, Gillian R Hayes, Courtney Heldreth, Michal Lahav, Jess Holbrook, and Lauren Wilcox. 2022. A Systematic Review and Thematic Analysis of Community-Collaborative Approaches to Computing Research. In *CHI Conference on Human Factors in Computing Systems*. 1–18.

[25] Claire Craig. 2020. Context is everything. , 139–141 pages.

[26] Helen Cunningham and Stephen Reay. 2019. Co-creating design for health in a city hospital: perceptions of value, opportunity and limitations from 'Designing Together'symposium. *Design for Health* 3, 1 (2019), 119–134.

[27] Fernando Delgado, Solon Barocas, and Karen Levy. 2022. An Uncommon Task: Participatory Design in Legal AI. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–23.

[28] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2021. Stakeholder Participation in AI: Beyond" Add Diverse Stakeholders and Stir". *arXiv preprint arXiv:2111.01122* (2021).

[29] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–23.

[30] Wesley Hanwen Deng, Manish Nagireddy, Michelle Seng Ah Lee, Jatinder Singh, Zhiwei Steven Wu, Kenneth Holstein, and Haiyi Zhu. 2022. Exploring How Machine Learning Practitioners (Try To) Use Fairness Toolkits. *arXiv preprint arXiv:2205.06922* (2022).

[31] Wesley Hanwen Deng, Nur Yildirim, Monica Chang, Motahhare Eslami, Kenneth Holstein, and Michael Madaio. 2023. Investigating Practices and Opportunities for Cross-functional Collaboration around AI Fairness in Industry Practice. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 705–716.

[32] Matt Dexter, Paul Atkinson, and Andrew Dearden. 2011. Health products; designed with, not for, end users. (2011).

[33] Aaron S Dietz, Peter J Pronovost, Pedro Alejandro Mendez-Tellez, Rhonda Wyskiel, Jill A Marsteller, David A Thompson, and Michael A Rosen. 2014. A systematic review of teamwork in the intensive care unit: what do we know about teamwork, team tasks, and improvement strategies? *Journal of critical care* 29, 6 (2014), 908–914.

[34] Graham Dove and Anne-Laure Fayard. 2020. Monsters, metaphors, and machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–17.

[35] 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*. 278–288.

[36] Graham Dove and Sara Jones. 2014. Using data to stimulate creative thinking in the design of new products and services. In *Proceedings of the 2014 conference on Designing interactive systems*. 443–452.

[37] Bar Eini-Porat, Ofra Amir, Danny Eytan, and Uri Shalit. 2022. Tell me something interesting: Clinical utility of machine learning prediction models in the ICU. *Journal of Biomedical Informatics* 132 (2022), 104107.

[38] Tatiana Ermakova, Julia Blume, Benjamin Fabian, Elena Fomenko, Marcus Berlin, and Manfred Hauswirth. 2021. Beyond the hype: why do data-driven projects fail?. In *Proceedings of the 54th Hawaii International Conference on System Sciences*. 5081.

[39] Michael Feffer, Michael Skirpan, Zachary Lipton, and Hoda Heidari. 2023. From Preference Elicitation to Participatory ML: A Critical Survey & Guidelines for Future Research. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 38–48.

[40] KJ Kevin Feng, Maxwell James Coppock, and David W McDonald. 2023. How Do UX Practitioners Communicate AI as a Design Material? Artifacts, Conceptions, and Propositions. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. 2263–2280.

[41] Jodi Forlizzi. 2018. Moving beyond user-centered design. *Interactions* 25, 5 (2018), 22–23.

[42] Astrid Galsgaard, Tom Doorschodt, Ann-Louise Holten, Felix Christoph Müller, Mikael Ploug Boesen, and Mario Maas. 2022. Artificial intelligence and multidisciplinary team meetings; a communication challenge for radiologists' sense of agency and position as spider in a web? *European Journal of Radiology* 155 (2022), 110231.

[43] 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.

[44] Pratik Ghosh, Karen L Posner, Stephanie L Hyland, Wil Van Cleve, Melissa Bristow, Dustin R Long, Konstantina Palla, Bala Nair, Christine Fong, Ronald Pauldine, et al. 2023. Framing Machine Learning Opportunities for Hypotension Prediction in Perioperative Care: A Socio-Technical Perspective. *ACM Transactions on Computer-Human Interaction* (2023).

[45] Jennifer C Ginestra, Rachel Kohn, Rebecca A Hubbard, Andrew Crane-Droesch, Scott D Halpern, Meeta Prasad Kerlin, and Gary E Weissman. 2022. Association of Unit Census with Delays in Antimicrobial Initiation among Ward Patients with Hospital-acquired Sepsis. *Annals of the American Thoracic Society* ja (2022).

[46] Fabien Girardin and Neal Lathia. 2017. When user experience designers partner with data scientists. In *2017 AAAI Spring Symposium Series*.

[47] Google. 2022. Machine Learning Guides: Text Classification. https://developers.google.com/machine-learning/guides/text-classification

[48] Divya Gopinath, Monica Agrawal, Luke Murray, Steven Horng, David Karger, and David Sontag. 2020. Fast, structured clinical documentation via contextual autocomplete. In *Machine Learning for Healthcare Conference*. PMLR, 842–870.

[49] Nielsen Norman Group. 2018. Using Prioritization Matrices to Inform UX Decisions. Retrieved September 14, 2022 from https://www.nngroup.com/articles/prioritization-matrices/

[50] Aaron Halfaker and R Stuart Geiger. 2020. Ores: Lowering barriers with participatory machine learning in wikipedia. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–37.

[51] Douglas D Heckathorn. 2011. Comment: Snowball versus respondent-driven sampling. *Sociological methodology* 41, 1 (2011), 355–366.

[52] Sean M Hickey and Al O Giwa. 2020. Mechanical Ventilation. In *StatPearls*. StatPearls Publishing, Treasure Island (FL).

[53] Tad Hirsch, Kritzia Merced, Shrikanth Narayanan, Zac E Imel, and David C Atkins. 2017. Designing contestability: Interaction design, machine learning, and mental health. In *Proceedings of the 2017 Conference on Designing Interactive Systems*. 95–99.

[54] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–16.

[55] Naja Holten Møller, Irina Shklovski, and Thomas T Hildebrandt. 2020. Shifting concepts of value: Designing algorithmic decision-support systems for public services. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*. 1–12.

[56] Aspen Hopkins, Fred Hohman, Luca Zappella, Xavier Suau Cuadros, and Dominik Moritz. 2023. Designing Data: Proactive Data Collection and Iteration for Machine Learning. *arXiv preprint arXiv:2301.10319* (2023).

[57] Solène Huynh-Dagher, Guillaume Lamé, Tu-Anh Duong, and Marija Jankovic. 2022. Design research in healthcare: a systematic literature review of key design journals. *Journal of Engineering Design* 33, 8-9 (2022), 522–544.

[58] IDEO. 2009. The Human-Centered Design Toolkit. https://www.designkit.org/

[59] Maia Jacobs, Jeffrey He, Melanie F. Pradier, Barbara Lam, Andrew C Ahn, Thomas H McCoy, Roy H Perlis, Finale Doshi-Velez, and Krzysztof Z Gajos. 2021. Designing AI for trust and collaboration in time-constrained medical decisions: a sociotechnical lens. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[60] Anniek Jansen and Sara Colombo. 2023. Mix & Match Machine Learning: An Ideation Toolkit to Design Machine Learning-Enabled Solutions. In *Proceedings of the Seventeenth International Conference on Tangible, Embedded, and Embodied Interaction*. 1–18.

[61] Stephanie Tulk Jesso, Aisling Kelliher, Harsh Sanghavi, Thomas Martin, and Sarah Henrickson Parker. 2022. Inclusion of clinicians in the development and evaluation of clinical artificial intelligence tools: a systematic literature review. *Frontiers in Psychology* 13 (2022).

[62] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data* 6, 1 (2019), 317.

[63] Mayur P Joshi, Ning Su, Robert D Austin, and Anand K Sundaram. 2021. Why so many data science projects fail to deliver. *MIT Sloan Management Review* (2021).

[64] Jeremy M Kahn, Christopher H Goss, Patrick J Heagerty, Andrew A Kramer, Chelsea R O'Brien, and Gordon D Rubenfeld. 2006. Hospital volume and the outcomes of mechanical ventilation. *New England Journal of Medicine* 355, 1 (2006), 41–50.

[65] Annika Kaltenhauser, Verena Rheinstädter, Andreas Butz, and Dieter P Wallach. 2020. " You Have to Piece the Puzzle Together" Implications for Designing Decision Support in Intensive Care. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. 1509–1522.

[66] Holtzblatt Karen and Jones Sandra. 2017. Contextual inquiry: A participatory technique for system design. In *Participatory design*. CRC Press, 177–210.

[67] Anna Kawakami, Venkatesh Sivaraman, Hao-Fei Cheng, Logan Stapleton, Yanghuidi Cheng, Diana Qing, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. 2022. Improving human-AI partnerships in child welfare: understanding worker practices, challenges, and desires for algorithmic decision support. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–18.

[68] Anna Kawakami, Venkatesh Sivaraman, Logan Stapleton, Hao-Fei Cheng, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. 2022. "Why Do I Care What's Similar?" Probing Challenges in AI-Assisted Child Welfare Decision-Making through Worker-AI Interface Design Concepts. In *Designing Interactive Systems Conference*. 454–470.

[69] Claire Kayacik, Sherol Chen, Signe Noerly, Jess Holbrook, Adam Roberts, and Douglas Eck. 2019. Identifying the intersections: User experience+ research scientist collaboration in a generative machine learning interface. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–8.

[70] Alireza Khanshan, Lars Giling, Panos Markopoulos, and Pieter Van Gorp. 2023. A Case Study of Data-Enabled Design for Cardiac Telemonitoring. In *Proceedings of the European Conference on Cognitive Ergonomics 2023*. 1–7.

[71] Andrew J King, Derek C Angus, Gregory F Cooper, Danielle L Mowery, Jennifer B Seaman, Kelly M Potter, Leigh A Bukowski, Ali Al-Khafaji, Scott R Gunn, and Jeremy M Kahn. 2023. A voice-based digital assistant for intelligent prompting of evidence-based practices during ICU rounds. *Journal of Biomedical Informatics* 146 (2023), 104483.

[72] Maaike Kleinsmann and Martijn Ten Bhömer. 2020. The (new) roles of prototypes during the co-development of digital product service systems. *International Journal of Design* 14, 1 (2020), 65–79.

[73] Rafal Kocielnik, Saleema Amershi, and Paul N Bennett. 2019. Will you accept an imperfect ai? exploring designs for adjusting end-user expectations of ai systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[74] Sean Kross and Philip Guo. 2021. Orienting, framing, bridging, magic, and counseling: How data scientists navigate the outer loop of client collaborations in industry and academia. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–28.

[75] Tzu-Sheng Kuo, Hong Shen, Jisoo Geum, Nev Jones, Jason I Hong, Haiyi Zhu, and Kenneth Holstein. 2023. Understanding Frontline Workers' and Unhoused Individuals' Perspectives on AI Used in Homeless Services. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.

[76] Michelle S Lam, Zixian Ma, Anne Li, Izequiel Freitas, Dakuo Wang, James A Landay, and Michael S Bernstein. 2023. Model Sketching: Centering Concepts in Early-Stage Machine Learning Model Design. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–24.

[77] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.

[78] Q Vera Liao, Hariharan Subramonyam, Jennifer Wang, and Jennifer Wortman Vaughan. 2023. Designerly understanding: Information needs for model transparency to support design ideation for AI-powered user experience. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–21.

[79] Yu-Wei Lin, Yuqian Zhou, Faraz Faghri, Michael J Shaw, and Roy H Campbell. 2019. Analysis and prediction of unplanned intensive care unit readmission

using recurrent neural networks with long short-term memory. *PloS one* 14, 7 (2019), e0218942.

[80] Houjiang Liu, Anubrata Das, Alexander Boltz, Didi Zhou, Daisy Pinaroc, Matthew Lease, and Min Kyung Lee. 2023. Human-centered NLP Fact-checking: Co-Designing with Fact-checkers using Matchmaking for AI. *arXiv preprint arXiv:2308.07213* (2023).

[81] Yaoli Mao, Dakuo Wang, Michael Muller, Kush R Varshney, Ioana Baldini, Casey Dugan, and Aleksandra Mojsilović. 2019. How data scientistswork together with domain experts in scientific collaborations: To find the right answer or to ask the right question? *Proceedings of the ACM on Human-Computer Interaction* 3, GROUP (2019), 1–23.

[82] Bella Martin, Bruce Hanington, and Bruce M Hanington. 2012. Universal methods of design: 100 ways to research complex problems. *Develop Innovative Ideas, and Design Effective Solutions* (2012), 12–13.

[83] Alessandro Morandi, Nathan E Brummel, and E Wesley Ely. 2011. Sedation, delirium and mechanical ventilation: the 'ABCDE'approach. *Current opinion in critical care* 17, 1 (2011), 43–49.

[84] Cecily Morrison, Edward Cutrell, Anupama Dhareshwar, Kevin Doherty, Anja Thieme, and Alex Taylor. 2017. Imagining artificial intelligence applications with people with visual disabilities using tactile ideation. In *Proceedings of the 19th international acm sigaccess conference on computers and accessibility*. 81–90.

[85] Camille Moussette and Richard Banks. 2010. Designing through making: exploring the simple haptic design space. In *Proceedings of the fifth international conference on Tangible, embedded, and embodied interaction*. 279–282.

[86] Dave Murray-Rust, Maria Luce Lupetti, Iohanna Nicenboim, and Wouter van der Hoog. 2023. Grasping AI: experiential exercises for designers. *AI & SOCIETY* (2023), 1–21.

[87] Dave Murray-Rust, Iohanna Nicenboim, and Dan Lockton. 2022. Metaphors for designers working with AI. (2022).

[88] Nadia Nahar, Shurui Zhou, Grace Lewis, and Christian Kästner. 2022. Collaboration Challenges in Building ML-Enabled Systems: Communication, Documentation, Engineering, and Process. *Organization* 1, 2 (2022), 3.

[89] Shamim Nemati, Andre Holder, Fereshteh Razmi, Matthew D Stanley, Gari D Clifford, and Timothy G Buchman. 2018. An interpretable machine learning model for accurate prediction of sepsis in the ICU. *Critical care medicine* 46, 4 (2018), 547.

[90] Renee Noortman, Peter Lovei, Mathias Funk, Eva Deckers, Stephan Wensveen, and Berry Eggen. 2022. Breaking up data-enabled design: expanding and scaling up for the clinical context. *AI EDAM* 36 (2022).

[91] Michael Oppermann and Tamara Munzner. 2020. Data-first visualization design studies. In *2020 IEEE Workshop on Evaluation and Beyond-Methodological Approaches to Visualization (BELIV)*. IEEE, 74–80.

[92] Tariq Osman Andersen, Francisco Nunes, Lauren Wilcox, Elizabeth Kaziunas, Stina Matthiesen, and Farah Magrabi. 2021. Realizing AI in healthcare: challenges appearing in the wild. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–5.

[93] Fatih Kursat Ozenc, Miso Kim, John Zimmerman, Stephen Oney, and Brad Myers. 2010. How to support designers in getting hold of the immaterial material of software. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2513–2522.

[94] Google PAIR. 2019. People + AI Guidebook. pair.withgoogle.com/guidebook

[95] Samir Passi and Solon Barocas. 2019. Problem formulation and fairness. In *Proceedings of the conference on fairness, accountability, and transparency*. 39–48.

[96] David Piorkowski, Soya Park, April Yi Wang, Dakuo Wang, Michael Muller, and Felix Portnoy. 2021. How ai developers overcome communication challenges in a multidisciplinary team: A case study. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–25.

[97] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. 2022. Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI. *arXiv preprint arXiv:2204.01075* (2022).

[98] Inioluwa Deborah Raji, I Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. 2022. The fallacy of AI functionality. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 959–972.

[99] Stephen Reay, Ivana Nakarada-Kordic, Nicola Kayes, Cassie Khoo, and Claire Craig. 2021. Initiate. collaborate: a design for health collaboration toolkit. *Design for Health* 5, 3 (2021), 294–312.

[100] Madhu C Reddy, Paul Dourish, and Wanda Pratt. 2001. Coordinating heterogeneous work: Information and representation in medical care. In *ECSCW 2001*. Springer, 239–258.

[101] Madhu C Reddy, Paul Dourish, and Wanda Pratt. 2006. Temporality in medical work: Time also matters. *Computer Supported Cooperative Work (CSCW)* 15, 1 (2006), 29–53.

[102] Johan Redström. 2005. On technology as material in design. *Design Philosophy Papers* 3, 2 (2005), 39–54.

[103] Eric Reis. 2011. The lean startup. *New York: Crown Business* 27 (2011), 2016–2020.

[104] Samantha Robertson and Niloufar Salehi. 2020. What If I Don't Like Any Of The Choices? The Limits of Preference Elicitation for Participatory Algorithm Design. *arXiv preprint arXiv:2007.06718* (2020).

[105] Negar Rostamzadeh, Diana Mincu, Subhrajit Roy, Andrew Smart, Lauren Wilcox, Mahima Pushkarna, Jessica Schrouff, Razvan Amironesei, Nyalleng Moorosi, and Katherine Heller. 2022. Healthsheet: development of a transparency artifact for health datasets. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1943–1961.

[106] Nithya Sambasivan and Rajesh Veeraraghavan. 2022. The Deskilling of Domain Expertise in AI Development. In *CHI Conference on Human Factors in Computing Systems*. 1–14.

[107] Martin G Seneviratne, Nigam H Shah, and Larry Chu. 2020. Bridging the implementation gap of machine learning in healthcare. *BMJ Innovations* 6, 2 (2020).

[108] Ben Shneiderman. 2020. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction* 36, 6 (2020), 495–504.

[109] Venkatesh Sivaraman, Leigh A Bukowski, Joel Levin, Jeremy M Kahn, and Adam Perer. 2023. Ignore, trust, or negotiate: understanding clinician acceptance of AI-based treatment recommendations in health care. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.

[110] J Spiegel, M McKenna, G Lakshman, and P Nordstrom. 2014. Amazon us patent anticipatory shipping. *Amazon Technologies Inc* 12 (2014).

[111] Preethi Srinivas, Anthony Faiola, and Gloria Mark. 2016. Designing guidelines for mobile health technology: Managing notification interruptions in the ICU. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 4502–4508.

[112] Logan Stapleton, Min Hun Lee, Diana Qing, Marya Wright, Alexandra Chouldechova, Ken Holstein, Zhiwei Steven Wu, and Haiyi Zhu. 2022. Imagining new futures beyond predictive systems in child welfare: A qualitative study with impacted stakeholders. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1162–1177.

[113] Simone Stumpf, Vidya Rajaram, Lida Li, Margaret Burnett, Thomas Dietterich, Erin Sullivan, Russell Drummond, and Jonathan Herlocker. 2007. Toward harnessing user feedback for machine learning. In *Proceedings of the 12th international conference on Intelligent user interfaces*. 82–91.

[114] Hariharan Subramonyam, Jane Im, Colleen Seifert, and Eytan Adar. 2022. Solving Separation-of-Concerns Problems in Collaborative Design of Human-AI Systems through Leaky Abstractions. In *CHI Conference on Human Factors in Computing Systems*. 1–21.

[115] Petra Sundström, Alex Taylor, Katja Grufberg, Niklas Wirström, Jordi Solsona Belenguer, and Marcus Lundén. 2011. Inspirational bits: towards a shared understanding of the digital material. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1561–1570.

[116] Harini Suresh, Nathan Hunt, Alistair Johnson, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. 2017. Clinical intervention prediction and understanding using deep networks. *arXiv preprint arXiv:1705.08498* (2017).

[117] Mohammad Tahaei, Marios Constantinides, Daniele Quercia, Sean Kennedy, Michael Muller, Simone Stumpf, Q Vera Liao, Ricardo Baeza-Yates, Lora Aroyo, Jess Holbrook, et al. 2023. Human-Centered Responsible Artificial Intelligence: Current & Future Trends. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–4.

[118] Mei Tan, Hansol Lee, Dakuo Wang, and Hariharan Subramonyam. 2023. Is a Seat at the Table Enough? Engaging Teachers and Students in Dataset Specification for ML in Education. *arXiv preprint arXiv:2311.05792* (2023).

[119] 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 Transactions on Computer-Human Interaction (TOCHI)* 27, 5 (2020), 1–53.

[120] Anja Thieme, Maryann Hanratty, Maria Lyons, Jorge Palacios, Rita Faia Marques, Cecily Morrison, and Gavin Doherty. 2023. Designing human-centered AI for mental health: Developing clinically relevant applications for online CBT treatment. *ACM Transactions on Computer-Human Interaction* 30, 2 (2023), 1–50.

[121] Anja Thieme, Aditya Nori, Marzyeh Ghassemi, Rishi Bommasani, Tariq Osman Andersen, and Ewa Luger. 2023. Foundation Models in Healthcare: Opportunities, Risks & Strategies Forward. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–4.

[122] Stephanie Valencia, Richard Cave, Krystal Kallarackal, Katie Seaver, Michael Terry, and Shaun K Kane. 2023. "The less I type, the better": How AI Language Models can Enhance or Impede Communication for AAC Users. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–14.

[123] Rama Adithya Varanasi and Nitesh Goyal. 2023. "It is currently hodgepodge": Examining AI/ML Practitioners' Challenges during Co-production of Responsible AI Values. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.

[124] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–14.

[125] Qiaosi Wang, Michael Madaio, Shaun Kane, Shivani Kapania, Michael Terry, and Lauren Wilcox. 2023. Designing Responsible AI: Adaptations of UX Practice

to Meet Responsible AI Challenges. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.

[126] Joyce Weiner. 2020. Why AI/data science projects fail: how to avoid project pitfalls. *Synthesis Lectures on Computation and Analytics* 1, 1 (2020), i–77.

[127] Jonathan West. 2020. Design in healthcare: The challenge of translation. *Design for Health* 4, 2 (2020), 252–269.

[128] Mikael Wiberg. 2014. Methodology for materiality: interaction design research through a material lens. *Personal and ubiquitous computing* 18, 3 (2014), 625–636.

[129] Lauren Wilcox, Robin Brewer, and Fernando Diaz. 2023. AI Consent Futures: A Case Study on Voice Data Collection with Clinicians. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–30.

[130] Lauren Wilcox, Jie Lu, Jennifer Lai, Steven Feiner, and Desmond Jordan. 2010. Physician-driven management of patient progress notes in an intensive care unit. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1879–1888.

[131] Qian Yang, Nikola Banovic, and John Zimmerman. 2018. Mapping machine learning advances from hci research to reveal starting places for design innovation. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–11.

[132] Qian Yang, Justin Cranshaw, Saleema Amershi, Shamsi T Iqbal, and Jaime Teevan. 2019. Sketching nlp: A case study of exploring the right things to design with language intelligence. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[133] Qian Yang, Yuexing Hao, Kexin Quan, Stephen Yang, Yiran Zhao, Volodymyr Kuleshov, and Fei Wang. 2023. Harnessing biomedical literature to calibrate clinicians' trust in AI decision support systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–14.

[134] Qian Yang, Alex Scuito, John Zimmerman, Jodi Forlizzi, and Aaron Steinfeld. 2018. Investigating how experienced UX designers effectively work with machine learning. In *Proceedings of the 2018 designing interactive systems conference*. 585–596.

[135] 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*. 1–13.

[136] 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*. 1–11.

[137] 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*. 4477–4488.

[138] Nur Yildirim, Alex Kass, Teresa Tung, Connor Upton, Donnacha Costello, Robert Giusti, Sinem Lacin, Sara Lovic, James M O'Neill, Rudi O'Reilly Meehan, et al. 2022. How Experienced Designers of Enterprise Applications Engage AI as a Design Material. In *CHI Conference on Human Factors in Computing Systems*. 1–13.

[139] Nur Yildirim, Changhoon Oh, Deniz Sayar, Kayla Brand, Supritha Challa, Violet Turri, Nina Crosby Walton, Anna Elise Wong, Jodi Forlizzi, James McCann, et al. 2023. Creating design resources to scaffold the ideation of AI concepts. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. 2326–2346.

[140] Nur Yildirim, Mahima Pushkarna, Nitesh Goyal, Martin Wattenberg, and Fernanda Viégas. 2023. Investigating How Practitioners Use Human-AI Guidelines: A Case Study on the People+ AI Guidebook. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–13.

[141] Nur Yildirim, Hannah Richardson, Maria T. Wetscherek, Junaid Bajwa, Joseph Jacob, Mark A. Pinnock, Stephen Harris, Daniel Coelho de Castro, Shruthi Bannur, Stephanie L. Hyland, Pratik Ghosh, Mercy Ranjit, Kenza Bouzid, Anton Schwaighofer, Fernando Pérez-García, Harshita Sharma, Ozan Oktay, Matthew Lungren, Javier Alvarez-Valle, Aditya Nori, and Anja Thieme. 2024. Multimodal Healthcare AI: Identifying and Designing Clinically Relevant Vision-Language Applications for Radiology. arXiv:2402.14252 [cs.HC]

[142] Nur Yildirim, John Zimmerman, and Sarah Preum. 2021. Technical Feasibility, Financial Viability, and Clinician Acceptance: On the Many Challenges to AI in Clinical Practice.. In *HUMAN@ AAAI Fall Symposium*.

[143] Nur Yildirim, Susanna Zlotnikov, Aradhana Venkat, Gursimran Chawla, Jennifer Kim, Leigh A. Bukowski, Jeremy M. Kahn, James McCann, and John Zimmerman. 2024. Investigating Why Clinicians Deviate from Standards of Care: Liberating Patients from Mechanical Ventilation in the ICU. arXiv:2402.13464 [cs.HC]

[144] Bowen Yu, Ye Yuan, Loren Terveen, Zhiwei Steven Wu, Jodi Forlizzi, and Haiyi Zhu. 2020. Keeping designers in the loop: Communicating inherent algorithmic trade-offs across multiple objectives. In *Proceedings of the 2020 ACM designing interactive systems conference*. 1245–1257.

[145] Kun-Hsing Yu, Andrew L Beam, and Isaac S Kohane. 2018. Artificial intelligence in healthcare. *Nature biomedical engineering* 2, 10 (2018), 719–731.

[146] Hubert Dariusz Zając, Natalia Rozalia Avlona, Finn Kensing, Tariq Osman Andersen, and Irina Shklovski. 2023. Ground Truth Or Dare: Factors Affecting The Creation Of Medical Datasets For Training AI. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 351–362.

[147] Hubert D Zając, Dana Li, Xiang Dai, Jonathan F Carlsen, Finn Kensing, and Tariq O Andersen. 2023. Clinician-facing AI in the Wild: Taking Stock of the Sociotechnical Challenges and Opportunities for HCI. *ACM Transactions on Computer-Human Interaction* 30, 2 (2023), 1–39.

[148] Sabah Zdanowska and Alex S Taylor. 2022. A study of UX practitioners roles in designing real-world, enterprise ML systems. In *CHI Conference on Human Factors in Computing Systems*. 1–15.

[149] Angie Zhang, Olympia Walker, Kaci Nguyen, Jiajun Dai, Anqing Chen, and Min Kyung Lee. 2023. Deliberating with AI: Improving Decision-Making for the Future through Participatory AI Design and Stakeholder Deliberation. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–32.

[150] John Zimmerman, Jodi Forlizzi, and Shelley Evenson. 2007. Research through design as a method for interaction design research in HCI. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 493–502.

[151] John Zimmerman, Changhoon Oh, Nur Yildirim, Alex Kass, Teresa Tung, and Jodi Forlizzi. 2020. UX designers pushing AI in the enterprise: a case for adaptive UIs. *Interactions* 28, 1 (2020), 72–77.

[152] John Zimmerman, Aaron Steinfeld, Anthony Tomasic, and Oscar J. Romero. 2022. Recentering Reframing as an RtD Contribution: The Case of Pivoting from Accessible Web Tables to a Conversational Internet. In *CHI Conference on Human Factors in Computing Systems*. 1–14.