Sociotechnical Challenge Modeling: A Design Method for Responsible AI in Healthcare and Social Welfare

Tyler Reinmund
tyler.reinmund@jesus.ox.ac.uk
University of Oxford
Oxford, UK

Lars Kunze
Lars.Kunze@uwe.ac.uk
University of the West of England
Bristol, UK

Marina Jirotka
marina.jirotka@cs.ox.ac.uk
University of Oxford
Oxford, UK

Abstract

We present Sociotechnical Challenge Modeling (STCM), a workshop-based design method to help healthcare and social welfare practitioners identify and address sociotechnical challenges in machine learning (ML) deployments. We evaluated STCM in a field experiment with two UK organizations, involving 26 practitioners including managers, data scientists, and frontline care professionals. The evaluation found that STCM cultivated a sociotechnical perspective by revealing interdependencies between ML tools and organizational practices. The physical cards stimulated exchange and experimentation, while the workshop fostered collaboration across disciplines. However, participants found predefined countermeasures too prescriptive, which prompted revisions to support more open-ended ideation. Our contributions are a novel design method for anticipating and mitigating sociotechnical challenges of ML in care settings, and an empirical evaluation of its perceived value and limitations. To support adoption and further research, all STCM materials, including editable card templates and worksheets, are available at: https://bit.ly/4plXkfi.

CCS Concepts

• Computing methodologies → Machine learning; **• Social and professional topics → Socio-technical systems**; **• Human-centered computing → Participatory design**; **• Applied computing**;

Keywords

Sociotechnical Challenges, Design Method, Evaluation, Responsible AI, Participatory Design, Healthcare, Social Welfare

1 Introduction

Health systems, government departments, and public welfare agencies across the world are applying artificial intelligence (AI) and machine learning (ML) on a multitude of care services [52, 75]. Machine learning tools[1] are being used to triage patients in preventive care programs [23, 61] and emergency departments [22, 45, 69], in radiology labs [42] and primary care clinics [78] to help with diagnoses, and child welfare agencies to contend with the high volume of referrals [14, 18, 37].

Concomitant with their growing deployment is the increasing awareness of the challenges associated with these tools: rigid and generic user workflows hinder local adaptation [22, 78]; fragmented public services produce incomplete and inaccurate data sets [10], proxy variables conceal – and reproduce – the consequences of structural discrimination [54], and fundamental assumptions of machine learning – for instance, that the trajectory of a current patient, service user, or child can be derived based on its similarity to those treated in the past [32, 37, 82] – contradict principles and values of professional practice. Many of these issues are what Chancellor refers to as "sociotechnical failures", or "failures embodied within the social and political world that the technical model inhabits" [13]. Facing such challenges, practitioners – front-line, technical, operational, and managerial professionals involved in the design, deployment, and use of ML tools – have expressed a need for methodological support in identifying and mitigating challenges that affect the ethical, effective, and efficient use of ML tools [33, 61, 85, 87].

Therefore, we created Sociotechnical Challenge Modeling (STCM): a design method[2] to support practitioners in their efforts to understand, anticipate, and address sociotechnical challenges on healthcare and social welfare machine learning deployments. STCM is delivered as a workshop composed of two activities: in the first activity, Challenge Prioritization, participants use a set of Sociotechnical Challenge Cards to identify and reason about known sociotechnical challenges; the second activity, Countermeasure Planning, requires participants to employ Countermeasure Worksheets to construct mitigation plans for high-priority sociotechnical challenges. We conducted a field experiment with two organizations to understand what value practitioners perceived in STCM and how it could be improved.

STCM was designed to achieve two primary and two secondary objectives. The primary objectives are to (1) support the identification of sociotechnical challenges within the context of a planned or ongoing machine learning project and (2) provide actionable guidance on how to address the identified challenges in resource-constrained organizational settings. In support of the two primary objectives, STCM aims to achieve two secondary objectives: (3) facilitate a shared understanding among the diverse stakeholders involved in the development, use, and maintenance of the ML tool and (4) deliver as an easy-to-use resource. In our field experiment, we evaluate STCM against these four objectives.

Our work makes the following two contributions to HCI: (1) a novel design method, known as Sociotechnical Challenge Modeling, that aids practitioners in identifying and addressing the sociotechnical challenges of machine learning in healthcare and social welfare; and (2) an empirical evaluation of STCM within two machine learning deployments, which provides evidence of the method's utility

[1]In this paper, we use the term ML tools to refer to computing systems that incorporate machine learning techniques. We focus on non-generative ML tools – predictive and analytical applications such as supervised classifiers and unsupervised clustering algorithms – rather than generative AI systems designed to produce novel outputs.

[2]By design method, we refer to a "combination of tools, toolkits, techniques and/or games that are strategically put together to address defined goals" [65, p. 196]. Tools refer to the material components of the method, a "collection of tools" is a toolkit, and a technique provides a description for how the tools are used [65, p. 196].

in practice and offers transferable design recommendations for future methods aimed at operationalizing Responsible AI (RAI) in the healthcare and social welfare sectors. To support adoption and further research, all STCM materials, including editable card templates and worksheets, are available at: https://bit.ly/4plXkfi.

The remainder of the paper is organized as follows. In Section 2, we situate STCM in the literature on RAI design methods and introduce the concept of sociotechnical challenges. Section 3 describes how STCM was created and outlines its components, while Sections 4 and 5 describe the methods and results of the field experiment, respectively. The paper ends with a discussion on the implications of our findings for the creation of design methods in HCI and a brief conclusion.

2 Background & Related Work

2.1 Design Methods for Responsible AI

Researchers across HCI, Fairness, Accountability, and Transparency (FAccT), and RAI have devised numerous methods to support the human-centered and responsible design and deployment of AI solutions. A distinction emerging in the literature is that between technical and normative interventions, or "making the thing right" and "making the right thing" [11, 68]. Technical design methods focus on ensuring that created ML tools adhere to accepted standards, principles, and practices of "good design" [1, 5, 15, 27, 30, 40, 44, 47]. In contrast, normative design methods aid researchers, designers, engineers, and professionals as they explore whether an ML tool should be designed at all. While more attention tends to be directed towards technical design methods [3], researchers have recently made substantial progress on normative methods [17, 60, 84]. One recent example exemplifies this line of scholarship. Kawakami et al. [36] created the *Situate AI Guidebook*, a protocol for public sector agencies to conduct deliberative workshops on deciding whether or not to deploy ML tools. The protocol consists of a series of questions that agency leadership, frontline workers, community advocates, and developers can ask to consider early on whether to progress with a given ML tool.

The distinction between technical and normative design methods is of course an artificial one, meant only to illustrate different areas of focus. For instance, it is imaginable that through the course of applying a technical design method, one could learn that the ML tool surfaces too many issues, prompting the discontinuation of the project. In this vein, there are several methods that help stakeholders consider the potential failures of an ML tool [33, 58]. Tang et al. [74] developed the *AI Failure Cards* to understand how impacted communities can mitigate AI failures in homelessness services. Similarly, the *AI Mismatch Approach* recently published by Saxena et al. [68] provides a framework for engineers and designers to anticipate algorithmic risk factors during the early stages of innovation.

Regardless of whether a method is technical or normative, it remains difficult for researchers to evaluate the design methods they create [4, 6, 7, 26]. Many evaluations are restricted to logical assessments and illustrative case studies [e.g., 35, 68], or rely on methods that do not reflect actual organizational contexts [e.g., 1, 9, 19, 33, 34]. Apart from these methodological limitations, a recent survey has identified that many evaluations focus on usability, rather than effectiveness, thereby limiting our understanding of how well these methods achieve their aims [6].

Our work builds on existing scholarship on design methods for RAI in several ways. First, whereas past toolkits have favored technical solutions to the social and ethical challenges raised by ML tools [80], STCM provides practitioners with a set of countermeasures across governance, workflow design, and stakeholder management, and, additionally, remains attentive to the role of power dynamics within organizational settings. Second, we evaluate STCM on two live ML deployments in the UK's healthcare and social welfare sectors. This evaluation moves beyond assessing only usability by considering STCM's performance across its objectives of identifying sociotechnical challenges, providing actionable guidance, facilitating shared understanding, and being easy-to-use. Finally, we focus on a novel set of challenges which prior RAI design methods have not yet addressed: the sociotechnical challenges of machine learning, which we introduce in the next section.

2.2 Sociotechnical Challenges of ML

Much of the previous research in HCI on design methods for AI and other forms of software focuses on translating academic knowledge into design practice. For instance, past work has created card-based methods to help designers create technologically-mediated nudges [12]; learn about technological acceptance theory [51], tangible learning games [20], and societal resilience [53], and the social impacts of ML [71]; or raising awareness about value-based design issues more broadly [24]. In this paper, we focus on a novel theoretical perspective that has not yet been integrated into a design method: the sociotechnical challenges of machine learning.

The concept of sociotechnical challenges has been variably defined across diverse writings. In one of its earliest formulations by Bostrom & Heinen [8] in 1977, sociotechnical challenges were conceived of as the outcome of a designer's fallible and partial understanding of people and organizations. This understanding has persisted until today, recapitulated in writings that refer to the place "where our abstractions and representations of ML do not match reality" [13, p. 84]. Others propose a more relational perspective, suggesting that sociotechnical challenges are due to interactions and interplays between technical and social factors [70, 86]. In earlier work, we discussed two limitations with the concept of sociotechnical challenges: it is imprecise and incomplete [62]. That is, there lacked clear criteria to distinguish sociotechnical issues from more general human-computer interaction difficulties, and existing explanations on the development of sociotechnical challenges contradicted one another and omitted crucial dimensions of analysis.

To strengthen the literature on the sociotechnical challenges of machine learning, we outlined the Sociotechnical Challenge Framework [62]. The framework draws from Orlikowski's Technologies-in-Practice [55–57] to define sociotechnical challenges as dissonances between the material properties of a machine learning tool and the rules and resources that are drawn on during recurrent social practices. It consists of three components: a multidimensional process model explaining how sociotechnical challenges emerge, a taxonomy of sociotechnical challenges, and a repository of countermeasures.

The first component of the Sociotechnical Challenge Framework is a model that describes three processes through which sociotechnical challenges may arise: (i) the translation of understandings, intentions, and values into material properties of an ML tool, (ii) technical requirements and infrastructural constraints for machine learning tools, and (iii) the integration of machine learning tools into care practices.

The second and third components of the framework are closely related. The taxonomy classifies eleven distinct types of sociotechnical challenges across the ML-enabled care pathway; the challenges are presented in Table 1. Meanwhile, the third component includes a repository of 50 countermeasures that practitioners can employ to mitigate the sociotechnical challenges. The countermeasures span various facets of an ML project: understanding the problem; designing the ML tool, workflow, and interface; continuously evaluating the program; training, aligning, and communicating with stakeholders; and establishing governance mechanisms for accountability and project management

In this paper, we integrated this theoretical perspective on sociotechnical challenges into an actionable design method for practitioners involved in the design, deployment, and use of ML tools in healthcare and social welfare.

3 Sociotechnical Challenge Modeling

This section details the iterative creation of STCM and documents its constituent materials.

3.1 Creating the STCM Design Method

This study represents the third project of a broader Research through Design (RtD) program. Building on prior phases of fieldwork and theoretical development [61–63], the present work focuses on the instantiation of these earlier insights into a concrete design method. We utilized RtD [89] because it privileges the generative, open-ended exploration required to translate abstract theory into material artifacts. To ensure methodological rigor and transparency in this process, we structured the inquiry across five phases: defining requirements, ideating solutions, developing a prototype, conducting a pilot study, and evaluating STCM through a field experiment. The first four phases are detailed in this section, while the method for the field evaluation is presented subsequently in Section 4.

3.1.1 Define Requirements. During discovery, we conducted fieldwork with a social welfare organization and a literature review. First, our fieldwork revealed that cross-functional groups of practitioners were implementing ML tools in resource constrained settings, and faced difficulties anticipating and mitigating various sociotechnical challenges [61]. The literature review informed key conceptual and practical decisions, such as the types of design methods that exist, their format and purpose, as well as considerations such as group size and meeting context [33, 38, 49, 50, 59, 65–67, 72, 79, 81, 83, 85, 87].

3.1.2 Ideate Solutions. While many theory-based design methods exist (see Section 2), their formats are tightly coupled with the specific theoretical concepts they aim to convey. This coupling makes adapting an existing method for a different theoretical framework infeasible. Given our focus on a previously unaddressed framework, we required a novel method, built specifically to operationalize its core concepts.

To ideate this new method, we conducted an adapted Crazy 8's exercise [28], generating eight distinct ideas in eight minutes. These ideas were then merged into three core concepts. We elaborated each concept according to its required materials, process, merits, limitations, and foundation in prior literature. Through this process, we selected STCM for prototyping. STCM was chosen because it met three key criteria: (1) it supports both analogue and digital formats, accommodating co-located and distributed teams; (2) it requires no specialized software or stand-alone applications; and (3) it closely aligned with the current design practices of the organizations engaged in our fieldwork.

3.1.3 Develop Prototype. A prototype of the design method was created in three design applications – *Figma*,[3] *Miro*,[4] and *Mural*[5] – to gather preliminary feedback. On 1 November 2023, the lead researcher presented the prototype to two groups of stakeholders: (i) practitioners from an adult social services department within a county council and (ii) leadership from an ML deployment in a regional team of the NHS.

3.1.4 Pilot Study. To support further iteration of the design method, a pilot study was conducted with participants from a Primary Care Network (PCN). The pilot study was conducted as a virtual workshop on 5 February 2024. Participants were split into 3 groups; members in each group were physically co-located.

At this point in the study, the design method consisted of three activities delivered through a *Mural* board:

(1) **Practice Mapping**: Participants constructed a visual representation of how the ML tool is used in practice on a *Mural* board.
(2) **Challenge Prioritization**: Participants reflected on whether and how the various types of sociotechnical challenges may arise using digital concept cards.
(3) **Countermeasure Planning**: Participants prioritized a set of predefined countermeasures associated with the identified sociotechnical challenges.

The experience of the pilot study led to significant revisions in the design method. First, the "Practice Mapping" activity was omitted to reduce the number of tasks required to be completed during the session, and allow more time for participants to focus on the concept of sociotechnical challenges and means to address them, the two primary objectives of the design method. Additionally, participants expressed frustration with the delivery of the design method as a virtual workshop. Therefore, two changes were made: the "Challenge Prioritization" activity was re-created in an analogue format with physical cards that can be used in face-to-face sessions and the "Countermeasure Planning" activity was implemented within an online spreadsheet application in which participants select relevant countermeasures from a predefined list on their personal computers.

[3]https://www.figma.com/
[4]https://miro.com/
[5]https://www.mural.co/

Table 1: Taxonomy of sociotechnical challenges of machine learning in healthcare and social welfare.

The ML-enabled Care Pathway		
Gathering Input Data	*Interpreting Model Outputs*	*Initiating a Care Action*
1. Data Inconsistencies 2. Poor Environmental Conditions 3. Data Restrictions	4. Hidden Predictors 5. Mismatched Objectives 6. Decision Point Disconnect	7. Insufficient Causal Information 8. Value Conflicts
	Multi-stage	
9. Disrupted Workflows	10. Limited Understanding	11. Additional Labor

3.2 STCM Materials

STCM is composed of several materials: 11 **Sociotechnical Challenge Cards**, a set of **Countermeasure Worksheets**, **Facilitator Guidance**, and **Workshop Templates**. All materials are available on the STCM website: https://bit.ly/4plXkfi.

3.2.1 Sociotechnical Challenge Cards. Each of the Sociotechnical Challenge Cards represents one of the 11 sociotechnical challenges defined in Section 2.2. The cards provide a device that enables practitioners to consider how specific sociotechnical challenges may affect their ML tool. Past HCI research on card-based design methods has found that cards support the accessibility of conceptual knowledge, facilitate collaboration by establishing a shared language and physical referents, and support participant engagement [12, 20, 51, 59].

As shown in Figure 1, the cards include six aspects, all of which are derived from the Sociotechnical Challenge Framework [62] and informed by previous research on card-based design methods [e.g., 12, 51, 53, 74]. Each card is assigned to one **Category**, according to the stage of the ML-enabled care pathway in which the sociotechnical challenge is likely to arise. On the front side of the cards, there is a **Title**, **Definition**, and an **Icon**. The Title and Definition provide succinct labels and descriptions for each concept to establish a shared vocabulary [20]; these aspects employ generalized language to help STCM users apply each challenge to a diversity of contexts. The Icon is a simple visual representation of the challenge to help participants distinguish between cards [51]. The back side of the cards includes an **Example** and **Questions**, two aspects which have been proposed as useful mechanisms for conveying complex theories to practitioners [73]. Inspired by past work which has found that design methods should provide domain-specific and realistic examples from a diversity of use cases [33, 85], the Examples provide scenarios of each sociotechnical challenge across various healthcare and social welfare services. Finally, the Questions are prompts intended to guide discussion during the workshop and stimulate reflection on each sociotechnical challenge.

3.2.2 Countermeasure Worksheet. To support the generation of resolutions to identified sociotechnical challenges, we developed a set of Countermeasure Worksheets illustrated in Figure 2. Drawing from past research, existing design methods, and participant feedback, the worksheet aims to balance convergent and divergent ideation.

In total, there are 11 Countermeasure Worksheets corresponding to the 11 sociotechnical challenges. The front of the worksheet includes an adapted Crazy 8's template [28]. Crazy 8's is a design method that requires participants to outline eight unique ideas in eight minutes; this technique challenges participants to engage in divergent thinking. On the back of the worksheet, a repository of predefined countermeasures for each sociotechnical challenge is provided. Participants synthesize the preselected countermeasures with those developed extemporaneously to create an actionable mitigation plan.

3.2.3 Facilitator Guidance & Workshop Templates. Beyond the two components directly utilized by participants, we developed three additional artifacts to enable others to employ STCM independently. These components include **Facilitator Guidance** and two **Workshop Templates**. The Facilitator Guidance delineates the tasks, considerations, and questions for those facilitating an STCM workshop; it is provided as a PDF file for easy dissemination and editable Word and markdown files to support local adaptation. The Workshop Templates include a spreadsheet and presentation in which participants can document the discussed sociotechnical challenges and the corresponding countermeasures.

4 Evaluating STCM through a Field Experiment

Following the creation of STCM, we conducted a field experiment to understand its perceived value among healthcare and social welfare practitioners. This section describes the study design.

4.1 Research Design

As the goal of the evaluation was to understand the perceived value of STCM for practitioners in healthcare and social welfare, it was important to evaluate the method with a sample of participants drawn from the target population of users. To further increase the validity of the results, it was also necessary to conduct the study in conditions that resembled the environments in which the design method will ultimately be used. From these considerations, a one-group field experiment [26, 48] was chosen as the most appropriate method as it offers the opportunity for researchers to evaluate interventions in naturalistic environments [76]. A one-group design excludes the use of a control group; no control was sought for both methodological and practical considerations: the research objective focuses on the perceived value the method provides to practitioners, rather than identifying causal relationships between use of the method and some outcome; and gaining access to research sites solely for use as a control proved infeasible.

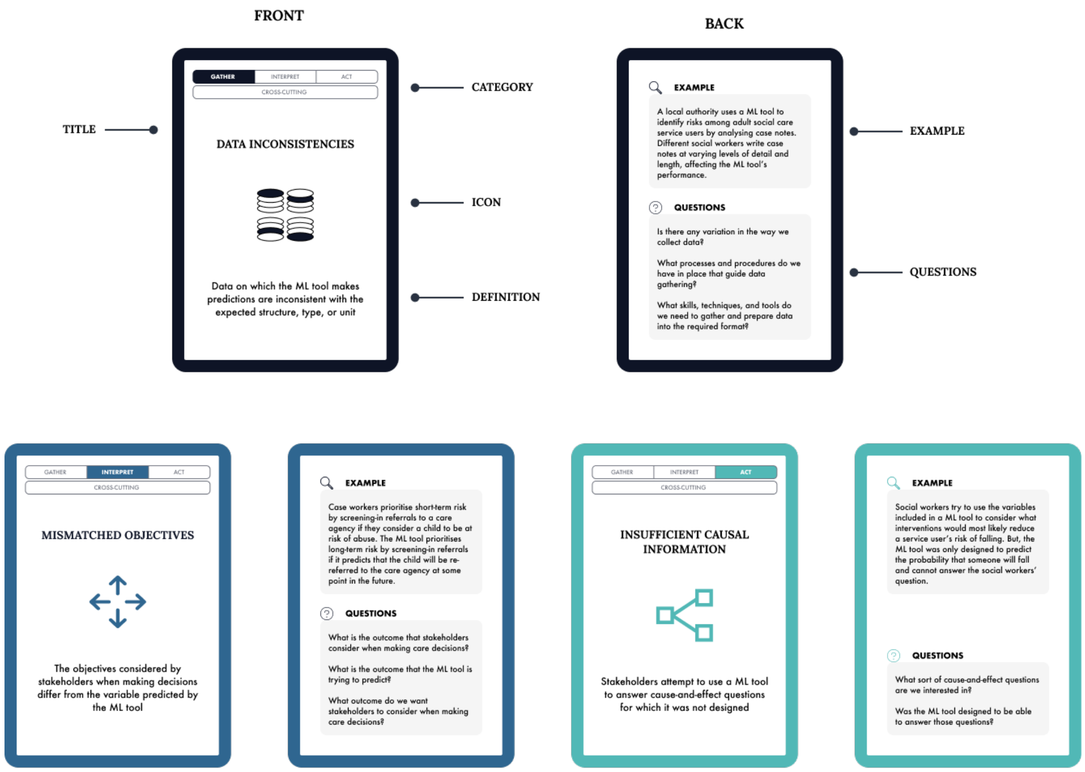
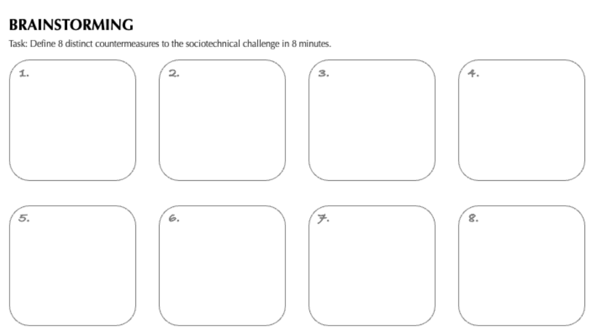
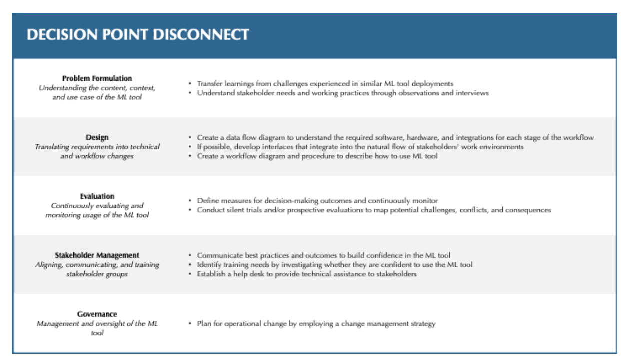


Figure 1: Overview of the Sociotechnical Challenge Cards



(a) Crazy 8's template



(b) Example countermeasures repository

Figure 2: Countermeasure worksheet

Two evaluation sessions were conducted with two healthcare and social welfare organizations in England. The first site was an adult social services department in the East of England. The second site was a PCN based in the South West of England. Seven participants from the PCN also participated in the pilot study. In this paper, the first site is referred to as **ASC01** and the second site is **HC01**.

4.2 Participants

Participant demographic information is provided in Table 2. Overall, 26 individuals participated in the field experiment. Pseudonyms represent the study, evaluation session, and participant number.

4.3 Protocol

The following steps comprise the evaluation process. All steps were followed at ASC01. The limited availability of participants imposed restrictions on the evaluation with HC01; as a result, step 5 was not performed and only 2 workshop attendees participated in an interview.

(1) **Planning**: Meetings were held with key stakeholders at each site to identify a time and location for the evaluation session, as well as determine who from each organization should participate.

Table 2: Overview of participants in field experiment. Note: Individuals who also participated in the pilot study are marked with an asterisk (*).

Site	ID	Role	Interview?	Site	ID	Role	Interview?
ASC01	S3-02-12	Prevention Manager	Y	HC01	S3-03-22	AI Fellow	Y
ASC01	S3-02-13	Director	Y	HC01	S3-03-23	AI Fellow	Y
ASC01	S3-02-14	Innovation Manager	Y	HC01	S3-03-24*	Chief Clinical Information Officer	N
ASC01	S3-02-15	Prevention Manager	Y	HC01	S3-03-25	General Practitioner	N
ASC01	S3-02-16	Data Manager	Y	HC01	S3-03-26*	Pharmacy Technician	N
ASC01	S3-02-17	Data Scientist	Y	HC01	S3-03-27	Pharmacy Technician	N
ASC01	S3-02-18	Project Manager	N	HC01	S3-03-28*	Pharmacy Technician	N
ASC01	S3-02-19	Prevention Manager	Y	HC01	S3-03-29*	Pharmacist	N
ASC01	S3-02-20	Data Analyst	Y	HC01	S3-03-30	Pharmacist	N
ASC01	S3-02-21	Data Analyst	Y	HC01	S3-03-31	Clinical Director	N
				HC01	S3-03-32	Health & Well-being Coach	N
				HC01	S3-03-33*	General Practitioner	N
				HC01	S3-03-34*	Occupational Therapist	N
				HC01	S3-03-35	Data Lead	N
				HC01	S3-03-36	PCN Manager	N
				HC01	S3-03-37*	Data Lead	N

(2) **Onboarding**: Participants received introductory emails outlining the purpose of the evaluation session. Participants at ASC01 received detailed instructions on the activities of the evaluation session. This information was not shared with participants from HC01 due to concerns among the site's key stakeholders that additional information may overburden participants.
(3) **Introduction**: The research team delivered an introductory presentation detailing the purpose, objectives, activities, and outcomes of the evaluation session.
(4) **Challenge Identification**: Participants employed "Sociotechnical Challenge Cards" to reflect on how various types of sociotechnical challenge may arise in the context of their own ML project. Participants prioritized each challenge according to their severity and probability. In ASC01, all eleven sociotechnical challenges were discussed. Due to the truncated evaluation session for HC01, only three sociotechnical challenges were discussed.
(5) **Countermeasure Planning**: Participants discussed plans to address the sociotechnical challenges categorized as "high" along the dimensions of severity and probability. Participants were grouped and assigned to a specific challenge where they generated a remediation plan by reviewing a list of predefined countermeasures.[6]
(6) **Interviews**: One-on-one interviews were conducted with the participants. The interview protocol was designed to elicit participant perceptions regarding the four evaluative criteria defined in Section 1.

[6]During the field experiment, participants employed a previous version of STCM in which the countermeasures were documented within an online spreadsheet application. Therefore, comments made by participants are in reference to this earlier version.

4.4 Data Collection

Qualitative data was collected through observations during the evaluation session and semi-structured interviews. An interview protocol was developed by adapting questionnaires and interview protocols employed in related work [48, 51]. Questions focused on whether participants perceived the method to be useful and easy to use, its impact on their understanding of their own ML project, their intentions to use the method in the future, and the clarity of the language employed in the design method materials. Interviews were conducted with 11 of the 26 participants. The interviews ranged in duration from 28 to 55 minutes, with a mean duration of 38 minutes and 33 seconds. With the consent of participants, the evaluation sessions and interviews were audio recorded and transcribed verbatim through Microsoft Teams. This study was granted approval by the University of Oxford's Computer Science Departmental Research Ethics Committee.

4.5 Data Analysis

We analyzed participants' perceptions on STCM using the Framework Method [25]. First, we defined the four objectives of STCM as deductive codes. After familiarization and anonymization, the lead researcher open-coded a sample of five transcripts to develop inductive codes. These codes either further specified the objectives (e.g., adding "Accessibility" under "Ease of Use") or captured emergent concepts like participant behavior. We grouped conceptually similar codes into an "analytical framework" [25] comprising 113 open codes within 6 themes and 26 categories. This framework was refined by recoding the initial transcripts to remove redundancies before being applied to the full dataset. Finally, we summarized data into a "framework matrix" (rows for participants, columns for categories) to facilitate interpretation and cross-participant. All research materials are available on the Open Science Framework.[7]

[7]https://doi.org/10.17605/OSF.IO/DRF5S

4.6 Positionality Statement

Our research team brings perspectives from HCI, Science and Technology Studies (STS), RAI, and Robotics. We are based in England and have engaged in fieldwork within the healthcare and social welfare sectors for four years. The lead author established long-term professional relationships with both partner organizations to negotiate access and facilitate the workshops. We acknowledge that these established relationships and our active role in facilitating the sessions likely influenced participant engagement and the workshop dynamics, a characteristic inherent to engaged RtD inquiries.

5 Findings

We present our findings in three parts: observations of participant behavior during the workshops, evaluations of STCM across its four objectives, and suggestions for improvement.

5.1 Participant Behavior Patterns

In the next two sections, we describe participants' behaviors during the two activities of STCM: (1) Challenge Prioritization and (2) Countermeasure Planning.

5.1.1 Behavior During Challenge Prioritization. During the Challenge Prioritization activity, several distinct behavior patterns among participants emerged. Participants employed the six aspects of the Sociotechnical Challenge Cards – the Title, Definition, Icon, Example, Questions, and Category – for a range of purposes. Based on feedback gathered during the interviews, the **participants perceived the example to be the most important aspect during their discussion**. And while employing STCM, **participants exploited the tangible nature of the cards to engage in two distinct collaborative processes: exchange and experimentation**.

Ten of the eleven participants interviewed used the Example during the activity. And in contrast to the other aspects, participants utilized the Example for multiple purposes. For instance, participants used the Example to align understanding between members of the group and reify the concepts presented on the cards. Further, participants stated that the inclusion of examples drawn from diverse domains provoked them to consider perspectives they may have otherwise overlooked, a finding that is elaborated in Section 5.2.

Finally, multiple participants emphasized the value of the Sociotechnical Challenge Cards' tangible format, with many contrasting the experience with hypothetical or past scenarios in which they performed similar activities on a computer. The physical nature of the cards enabled participants to engage in two collaborative processes: exchange and experimentation. Exchange involved participants independently reading the card, passing it to their colleague, reflecting on it, and repeating.

As the participants began prioritizing the challenges along the likelihood by severity scale, they engaged in experimentation. They would envision placing a card in one position on the scale, discuss the implications of its placement with their colleagues, and adjust the placement accordingly. During this time, they would explore potential disagreements between different interpretations of the challenges and classifications of their severity and likelihood.

5.1.2 Behavior During Countermeasure Planning. As with Challenge Prioritization, several distinct behaviors occurred during the second activity. During the field experiment, participants employed an earlier version of the design method in which participants only selected countermeasures from a predefined list; the list was distributed through an online spreadsheet application. These participants experienced technical issues with this delivery method, leading to the emergence of several workarounds. Further, **most participants relied on the set of predefined countermeasures in a prescriptive manner but contextualized them to better align with the nuances of their organizational context**; this finding is discussed in Section 5.2.2.

Participants experienced issues when attempting to use the online spreadsheet of countermeasures. The spreadsheet included a feature for participants to filter the list of countermeasures by those that were relevant to their assigned sociotechnical challenge. Several participants were unable to successfully utilize this functionality, attributing this difficulty to their lack of familiarity with the application. In response to this technical issue, participants adopted several workarounds: one group downloaded the spreadsheet locally so it could be opened in an application with which the participants were more familiar; another group avoided use of the filtering functionality altogether and manually examined each countermeasure to determine whether it was relevant to their assigned challenge.

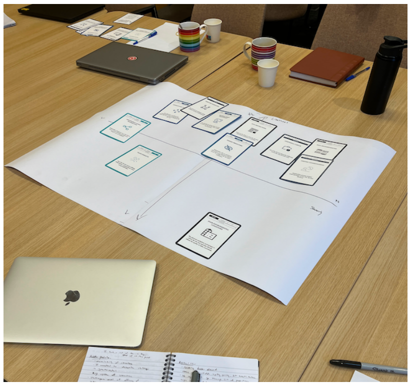
5.2 Perceived Value

This section presents participants' perceptions of the value provided by STCM. Results are organized according to STCM's four objectives: (1) identify sociotechnical challenges, (2) provide actionable guidance, (3) facilitate shared understanding, and (4) ease of use. Each objective is further divided into two to three dimensions; as described in Section 4.5, these dimensions were defined inductively through analysis of workshop observations and interview transcripts.

5.2.1 OB1: Identify Sociotechnical Challenges. STCM's first objective is to support the identification of sociotechnical challenges. Participants' evaluations focused on three dimensions: novelty, systematicity, and validity.

For **novelty** (whether STCM reveals new considerations), feedback was mostly positive. **STCM prompted participants to consider the relationship between their ML tool and its organizational and environmental context and revealed novel challenges they had not previously considered.**

STCM catalyzed a shift from a myopic "technical" view to one considering complex social and technical interdependencies. Participants said STCM enabled them to "*think outside the box*" [S3-02-12] and "*think about... [the ML tool] in a different way*" [S3-02-13]. Some, especially technical participants (data analysts, database administrators), shared that STCM broadened their perspective on machine learning to appreciate the interdependencies between their technical concerns (data quality, efficiency, predictive accuracy) and the values of the many stakeholders. This quote illustrates the change:



(a) Prioritized Sociotechnical Challenge Cards

(b) Participants discussing countermeasures to address high-priority sociotechnical challenges

Figure 3: Practitioners from ASC01 participating in an STCM workshop

> "*Hey, at this point I think it's not about the technology anymore, right?...[STCM] helped me get a broader understanding of how to approach machine learning because, for a person who is in technology might just continue to think, 'oh, it's just the tech, just the tech, just the tech' and then getting to see that, 'hey, it's not just the tech', you need to begin to see from the different perspectives, the different views of people, the stakeholders.*" [S3-02-20]

Others found the Sociotechnical Challenge Cards' challenges novel in themselves; for instance, "Poor Environmental Conditions" was a challenge participants had not previously considered.

The second dimension, **systematicity**, asks whether STCM facilitates a systematic approach to anticipating challenges. Unlike novelty, participants were more equivocal on this dimension.

First, the **Sociotechnical Challenge Cards provided a "*systematic framework*" that focused discussion.** Participants compared this to other workshops where they "*start[ed] from scratch*" [S3-02-15]:

> "*Often, I'll go to workshops where it's almost like a blank piece of paper. So it might have been, you know, we're here to talk about the wider ramifications of machine learning for the council...but I think having challenges already identified was super helpful.*" [S3-02-13]

Participants also found **the cards' content stimulated comprehensive and analytical reflection.** For instance, S3-02-15 said the categorization prevented getting "*bogged down*" in data gathering and overlooking other challenges. The cards' diverse examples also initiated a process of translating abstract challenges into practical terms for their specific project. Conversely, negative perceptions centered on the usefulness of Challenge Prioritization as a standalone activity and the rating scale employed. First, participants found **the rating scale arbitrary and subject to multiple interpretations, causing struggles in reaching agreement.** Second, HC01 participants, who did not complete Countermeasure Planning, **questioned the utility of Challenge Prioritization as an independent exercise.** Participants stated it is insufficient to only discuss challenges; this must be followed by countermeasures to prevent frustration:

> "*I do think that the second part is the part that makes it useful...I don't think, if I'm being honest, the first activity by itself is that useful unless something happens afterwards...if that was a stand alone exercise by itself, I feel like it just kind of got everyone riled up and then went away.*" [S3-03-22]

The final dimension was validity: the relevance of the sociotechnical challenges to participants' ML tools, assessing face validity. Participants **positively evaluated the 11 sociotechnical challenges' relevance to various ML applications and the appropriateness of the ML-enabled care pathway as a categorization scheme.** However, participants **may have misunderstood the ML-enabled care pathway.** S3-02-15 described the challenges' applicability to her project and other ML uses, relating this validity to a high intention for future use.

> "*These are challenges that are seen across the board...these are things that are considered in other areas as well. And so for us, if we go to do wider things around prevention and using machine learning, these are things that are common challenges.*" [S3-02-15]

Another participant related several sociotechnical challenges to ongoing conflicts in her ML project. This intuitive understanding and ability to connect abstract concepts to practical work supports the challenges' validity. Finally, participants commented on the ML-enabled care pathway's alignment with common ML workflows. While described as "*apt*", some participants seemed to misunderstand it as the "*typical process for developing an ML system*" [S3-02-17], rather than the human-user interaction process (Section 2.2).

5.2.2 OB2: Provide Actionable Guidance. STCM's second objective is to provide actionable guidance for addressing sociotechnical challenges in resource-constrained settings. Participants' evaluations focused on three dimensions: generative, pragmatic, and proactive.

Regarding the **generative** capacity (whether STCM produces countermeasures), participants gave mostly negative and ambivalent feedback on the earlier version. While some found predefined countermeasures simple ("*half the thinking is done for you*" [S3-02-17]), allowing them to contextualize for their project [S3-02-19], most from ASC01 stated that **the process of selecting predefined countermeasures constrained their ability to consider novel ideas.** This process, one participant noted, "*stifled any kind of wider creative thinking*" [S3-02-13]. **In response, STCM was revised to enable both divergent brainstorming and convergent selection.** The revised Countermeasure Planning activity was shown to 2 HC01 participants, who anticipated the generative process would be useful, with S3-03-23 calling it an "*options analysis*".

For the **pragmatic** dimension (creating feasible plans), evaluations diverged between the two sites. ASC01 shared positive feedback, while HC01 were uncertain. ASC01 participants **evidenced feasibility by reporting that they have begun to incorporate countermeasures within project planning.** S3-02-13, a director, wanted to "*take the outcomes...and use them as part of the...future of the roll out,*" while two other participants S3-02-15 and S3-02-17 later reported discussing the countermeasures in separate meetings: "*it wasn't just a workshop that we did and then forgot about it.*" [S3-02-17] In contrast, HC01 participants suggested frontline staff may not have the resources or capacity to implement the proposed countermeasures.

Finally, the **proactive** dimension considered whether STCM provoked planning. Participants expressed that **STCM caused them to consider opportunities afforded by the adoption of ML and challenged them to anticipate how sociotechnical challenges may arise in future ML projects.** During the ASC01 workshop, discussion of challenges unexpectedly gravitated towards opportunities. For example, the "Mismatched Objectives" challenge led to a discussion on improving decision-making consistency:

> **S3-02-17:** "*We were talking about...10 different workers coming up with 10 different ideas...It would provide a consistency...I think it would give a consistent approach.*"
> **S3-02-13:** "*I like that a lot. I think if you're doing population-level prevention...that consistency would be quite helpful.*"

Reflecting during the interview, S3-02-13 described STCM as "*a methodology*" and "*a bit like...a risk management approach, but...It's about the opportunities that might come out of something as well.*" Participants also felt STCM instigated them to consider challenges in future ML projects, with S3-02-17 and S3-02-15 describing the importance of "*mak[ing] sure that all of these challenges are appropriately scoped*".

5.2.3 OB3: Facilitate Shared Understanding. STCM's third objective is to facilitate shared understanding among diverse stakeholders. Participants' evaluations focused on two dimensions: expectations and information sharing.

For **expectations**, participants **valued the opportunity to discuss practical details of the ML tool with senior leaders and learn about conflicting judgments of the ML tool.** ASC01 participants reported STCM enabled discussing technical aspects with senior leaders, who usually lack visibility into these concerns. For instance, S3-02-14 noted: "*It was great to have [the director] there so that he could actually see what this means in practice*". The director corroborated this, indicating a new appreciation for the program's technical and operational complexity:

> "*In the last week or so...we've been giving much more thought to the data sharing that will be required. If we want to roll this out...there will be a whole new level of data sharing infrastructure...that maybe we hadn't fully appreciated the complexity of.*" [S3-02-13]

Meanwhile, HC01 participants responsible for deployment were surprised to learn about conflicting evaluations and end-user frustrations, which they previously had little exposure to. S3-03-23 illustrated his realization of misaligned expectations:

> "*So it was the first time I'd had a sit down with a clinical team who were actively using the tool... I was a little surprised when the workshop turned more towards some criticisms of the of the tool and how they use it and their confidence in it.*" [S3-03-23]

Regarding **information sharing**, two findings emerged. First, **participants with limited experience with a given ML tool reported that STCM helped improve their understanding of it; this effect was absent for those with extensive experience.** Second, participants **valued learning about the challenges faced by professionals in other roles.** The first finding is expected: participants with limited experience (e.g., a recently on-boarded data analyst [S3-02-21]) learned more than those with extensive experience (e.g., the data scientist [S3-02-17]):

> "*I would say I had zero knowledge about the [ML tool] before the workshop, and I would say I have at least 30 to 40%.*" [S3-02-21]
> "*I don't think the workshop specifically affected my knowledge about the [ML tool], given that I'm more technically involved with it.*" [S3-02-17]

While information gain varied, learning about others' challenges was widely shared. Participants valued "*learning about what other people were saying, thinking*" [S3-02-14] to identify "*different things that you might not have considered*" [S3-02-15]. For instance, a commissioner gained a "*much better understanding*" of data scientists' challenges, while a data analyst reported a newfound appreciation for social workers' pressures:

> "*When [S3-02-19] mentioned that social workers are often swarmed with... reports... I never thought about that honestly...when you get to see from the social worker's perspective...these are the challenges they face.*" [S3-02-20]

5.2.4 OB4: Ease of Use. STCM's final objective is to be an easy-to-use resource. Participants' evaluations focused on three dimensions: accessibility, comprehensibility, and workload.

For **accessibility**, most participants gave positive responses: **they stated that the materials accommodate collaboration between diverse stakeholders, the process is simple, and the overall experience is enjoyable and game-like.** Conversely, negative comments targeted **the inaccessibility of some language and technical difficulties with online spreadsheet applications.** Participants suggested the cards' presentation and content (e.g., examples, tangibility) help people with different expertise and working styles engage. The physical cards were also valued for being persistent and easily retrievable, unlike a screen [S3-02-13].

> "*I think the fact [the cards] were physically printed...I would absolutely stick with that. It made such a difference...you didn't have to constantly be looking up at the screen.*" [S3-02-13]

The cards' presentation reminded participants of a game [Field notes, 12 November 2024], making the experience enjoyable. Importantly, participants perceived the process as "*quite simple to do*" [S3-02-13], with little reliance on an expert facilitator:

> "*I think it'd be fairly easy for someone to pick this up...and kind of run with it without having you be there, which I don't mean in an offensive way, I mean in a...complimentary way.*" [S3-03-22]

While overwhelmingly positive, accessibility issues were noted. Some language was found complex (as elaborated next), and technical difficulties with the online spreadsheet application, along with feedback from Section 5.2.2, led to the redesign of the Countermeasure Planning activity.

Regarding **comprehensibility** (understanding concepts and language), **participants provided conflicting evaluations, with individual perspectives varying along professional role.** Two patterns emerged: participants in technical roles (database administrators, data scientists) found the language clear, while non-technical participants found it "*quite complex*" [S3-02-13] and "*abstract*" [S3-02-14], potentially impacting accessibility.

Finally, for **workload**, participants focused on the quantity of information and workshop pace, offering diverging perspectives on whether the total number of Sociotechnical Challenge Cards is tractable.

> "*There was enough to cover a whole area...but actually there wasn't too many.*" [S3-02-16]
> "*Part of me thinks we should even just like focus on just one of the cards...so you can get from start to finish.*" [S3-03-22]

This contrast is partially explained by workshop differences. ASC01 (S3-02-16) covered 11 challenges in 105 minutes (9.54 min. / challenge). HC01 (S3-03-22) discussed 3 challenges in 47 minutes (15.67 min. / challenge). The facilitator found HC01 more difficult to moderate, partially due to differences in volubility between participants; at that pace, 11 challenges would have taken nearly three hours, proving burdensome for HC01 participants.

5.3 Recommendations for Improvement

This section reports on recommendations for improving STCM. Each recommendation is based on feedback gathered from participants and our experience employing STCM.

Establish a Formal Prioritization Scale: The likelihood by severity scale used to prioritize sociotechnical challenges suffered from several limitations. During the workshop with ASC01, a prioritization scale was created on a large sheet of paper: two intersecting axes were drawn with the maximum and minimums labeled as "high" and "low", respectively (see Figure 3a). As a result, participants were unsure as to whether the scale was binary or graded. This led participants to classify each sociotechnical challenge as either "high" or "low" severity and likelihood, limiting the distance between each challenge on the prioritization scale. Relatedly, another participant stated that people may have different interpretations of the meaning of severity: "*Some people might have viewed it internally from a project success perspective. Some people have viewed it from a company reputation perspective*" [S3-02-17]. Participants may have relied on different implicit definitions of severity when prioritizing challenges, potentially confounding the ratings.

STCM does not prescribe a rating scale so that practitioners from diverse organizations can easily adapt the design method to align with their internal procedures. Therefore, to adhere to this requirement while addressing the concerns identified by participants, the prioritization of sociotechnical challenges can be improved by dedicating time prior to employing STCM by defining and agreeing upon a graded scale.

Resolve Misinterpretations of Sociotechnical Challenges: While some participants highlighted misinterpretations of the prioritization scale as an issue, others pointed to conflicting or deficient understandings of the sociotechnical challenges themselves. Participants suggested that it would be useful to dedicate time prior to the first activity for the facilitator to read out each Sociotechnical Challenge Card and address any questions or uncertainties among participants. From the perspective of participants, this would ensure that the time spent in the group discussion would be focused on the appropriate topics.

During the workshops, we refrained from reading the cards out to participants for evaluation purposes. That is, given that we were evaluating the comprehensibility of the cards, we attempted to limit any influence we could have on the perceived clarity of the cards by, for example, inadvertently resolving ambiguities. Outside of an evaluation setting, it is important to resolve uncertainties surrounding the sociotechnical challenge concepts prior to discussion to ensure that participants have a clear understanding of the topics they are meant to cover during the workshop activities.

Facilitate Cross-organizational Countermeasure Planning: As found in previous work, the design, deployment, and use of an ML tool brings together the contributions of individuals from multiple cross-sector organizations: software providers, local authorities, consulting firms, health services, and volunteer organizations [39, 46, 61]. This finding has implications for the planning of countermeasures, as identified by several participants. For example, as reported in Section 5.2.2, participants shared that it would be difficult for certain organizations to implement the identified countermeasures independently; rather, they would require the assistance of partners, such as regional NHS teams. Ensuring that such organizations participate in Countermeasure Planning and are accountable for implementing the mitigation plans presents a significant complexity for teams implementing ML tools, a reality not lost on participants:

> "*There are some challenges that we can't face on our own. A lot of these things have to be considerably collaborative. And if another organization has a lot of say and sway around an issue that matters to you, but they won't benefit from it, then how you manage that is really interesting.*" [S3-02-14]

Extending the Use of Examples: Throughout the evaluation, it became clear that participants placed a high value on the use of examples as a technique to reify the sociotechnical challenges and as referents to employ during the group discussion. When asked for recommendations to improve the STCM materials, participants stated that the use of examples should be extended in terms of their diversity, content, and context.

As discussed in Section 5.2.1, multiple participants valued the use of examples from adjacent domains; these individuals perceived that diverse examples stimulated richer discussion than would have occurred were all the examples specific to their context. Given their positive response to the examples from adjacent domains, some participants suggested diversifying the examples even further.

Additionally, one participant reported adjusting the content of the examples so that they include details on the potential consequences of sociotechnical challenges:

> "*You might highlight examples of companies or case studies that didn't necessarily mitigate these challenges, and something happened to them to give you further context...where people might not appreciate the impact of a particular challenge.*" [S3-02-17]

Finally, participants recommended adding examples to the countermeasures, in addition to the challenges. As noted by several participants, the predefined countermeasures are generic and high-level. As a result, some participants had difficulty understanding what a given countermeasure would look like in practice. Therefore, participants argued that including examples of countermeasures would help make them more tangible.

6 Discussion

In this section, we provide six implications for research on the creation of design methods across three domains, and outline opportunities for future research.

6.1 Collaborating through Physical Materials

A central finding of this study is that **participants exploited the tangibility of the Sociotechnical Challenge Cards to engage in the collaborative practices of exchange and experimentation**. During their discussions on the sociotechnical challenges, participants passed the cards back and forth between one another: they would read the front side of the card, turn it over, read the back, hand it to their partner, and pause to reflect on the challenge before initiating a conversation with their colleague. After an initial discussion, participants returned to the larger group and experimented with different configurations of the cards along the prioritization scale. They relied on the visual feedback of the cards' placement to test the implications of ranking a challenge as either higher or lower priority.

Participants recognized the utility of the physical cards for collaboration with many explicitly stating a preference for analogue design methods. Among the reasons given, participants placed importance on aspects both directly related to the design method and those that are associated with in-person interaction more generally: the ease with which the physical cards can be referred to during discussion and the ability to manipulate them; the capacity to maintain eye contact with other participants and the opportunity to engage in informal conversation. In the context of healthcare and social welfare, this mode of interaction aligns with the practice in the UK and elsewhere of multidisciplinary team meetings[2, 21, 43], where diverse professionals gather to discuss complex cases; because these sectors prioritize in-person collaboration, the design of the Sociotechnical Challenge Cards integrates with existing practices without the need to establish new working patterns.

These findings are particularly salient given trends in distributed work and the availability of collaboration software such as virtual conferencing (e.g., *Teams*, *Zoom*) and workspace applications (e.g., *Miro*, *Mural*, *Figma*). Further, some HCI scholars have suggested that research should focus on implementing design methods into virtual workspaces, as they may reduce barriers to adoption [85]. Virtual conferencing and workspace applications when used in conjunction offer features that attempt to emulate characteristics of in-person collaboration [cf. 41]: the verbal cues, gestures, and eye gaze necessary for turn-taking are afforded by video and audio transmission; the awkwardness of initiating communication is attenuated by "break out rooms"; the ability to resolve misunderstandings and disagreements is supported by the mutability of workspace applications; and task awareness is maintained through the passive display of collaborators' progress.

Despite the availability of features that aim to afford central actions of collaboration, participants – most, if not all, of whom have experience with virtual conferencing and workspace applications – still expressed a clear preference for physical design methods to be used during face-to-face interaction. These participants seem to implicitly make the claim that virtual conferencing and workspace applications do not offer a substitute for in-person collaboration: just an imperfect approximation of it.

Therefore, based on the findings presented here, we offer implications that diverge from what has been proposed previously in the literature:

> **Implication 1:** When the objective is to stimulate participation and collaboration, HCI researchers should, whenever possible, focus on developing physical design methods that can be employed within in-person interaction.

> **Implication 2:** To help reduce barriers to entry, researchers should make these resources freely available online in printable formats. For resource-constrained organizations, this removes friction; as observed in our pilot, access to proprietary workspace applications may be restricted or delayed by procurement processes, making low-fidelity, printable alternatives essential for immediate adoption.

6.2 Translating Theory into Design Methods

Methodological texts in the computing design disciplines often present the researcher as one who must strike a balance between two opposing objectives [e.g., 31, 88]. As academics, we aim to produce knowledge that stands up to the stringent requirements of scientific rigor; as members of a broader society whose research is funded by taxpayers and businesses, our work must be directed at problems people consider important and relevant. This dualism between rigor and relevance persists in the literature on design methods: researchers struggle to balance the complexity and granularity of theory with designers' needs for clarity, simplicity, and practicality [16, 73, 77]. Given the perennial and widespread nature of this problem, what does this study teach us about techniques to translate theory into design methods?

Through the design and evaluation of STCM, we found that **participants considered examples to be the most effective mechanism through which to communicate theoretical information**. Participants reported that the examples of sociotechnical challenges drawn from actual case studies made the concepts more tangible and served as referents towards which they could orient their discussion.

From the evaluation, we also learned that the mere presence of examples is insufficient for communicating theoretical concepts: participants value examples which are diverse and contain contextually relevant information. As discussed earlier, participants found the examples from domains adjacent to their own to be particularly useful in stimulating discussion. Examples from a novel domain allowed participants to abstract from the specifics of their particular context. And apart from broadening the diversity of examples, participants provided feedback on the information conveyed in the examples themselves. In STCM, the examples were devised to contain the following details: the setting (healthcare vs. social welfare), the intended use of the ML tool, the impacted stakeholder, and a manifestation of the relevant sociotechnical challenge. One piece of information that was excluded were details on what would occur if the sociotechnical challenge was left unmitigated. This omission was identified by one participant, which he provided as a suggestion to improve STCM.

Implication 3: Theory-based design methods should employ examples as an effective technique to reify concepts for practitioners.

Implication 4: Examples drawn from a diversity of adjacent domains can help stimulate discussion by provoking practitioners to think beyond the particularities of their own context.

Implication 5: When deciding what content to include within an example, HCI researchers should adopt an iterative approach, testing multiple variations with prospective users to understand what information they deem contextually relevant, as it is likely to vary on a case-by-case basis: information about consequences seems relevant for cards that aim to provoke discussion on actions, challenges, and risks, whereas such detail may be less appropriate for cards that aim to stimulate reflection on concepts such as values or technology acceptance factors [e.g., 51].

6.3 Bounded Divergence in Solution Ideation

The previous section illustrates an instance in which predictions about practitioner behavior with respect to the use of design methods proved correct: examples are a useful technique to help practitioners grapple with abstract theoretical concepts [73, 85]. In contrast, this section highlights a situation in which an initial prediction did not hold up. When creating STCM, we initially designed the Countermeasure Planning activity to consist of practitioners selecting interventions from a predefined repository; the intention was to allay any frustration which could arise from "bluesky thinking": unconstrained ideation that generates infeasible solutions [29, 74].[8] Yet, during the interviews, it became clear that participants found this approach constraining: they believed that this technique limited their ability to think creatively and move beyond what was already proposed.

Interestingly, despite participants criticizing the prescriptive nature of the Countermeasure Planning activity, some of the same individuals positively evaluated the structure provided within the Challenge Identification activity. They reported that "*having challenges already identified was super helpful*", comparing the experience to past workshops in which they began with "*a blank piece of paper*" [S3-02-13]. As this comparison illustrates, participants do not want to follow a rigid procedure, nor do they want to be left with only a blank canvas.

The problem for creators of design methods, then, is how to appropriately balance research-based knowledge on potential solutions with divergent ideation. That is, **how can we share with practitioners predefined interventions that are rooted in research while not curtailing their ability to think creatively?** To address this problem, we redesigned the Countermeasure Planning activity to comprise both divergent and convergent ideation phases: participants begin by employing the Crazy 8's method to generate 8 unique design ideas and then review a set of predefined countermeasures to produce a mitigation plan. While we did not have the opportunity to employ the revised activity in a workshop, preliminary feedback from participants indicates that it has merit, but will require further research.

Implication 6: When creating design methods that will be used by practitioners to define solutions to a problem, HCI researchers should devise techniques to support "bounded divergence". As per prior work, some individuals may consider unconstrained ideation to be frustrating; but, equally, others may find the controlled process of building on solutions created by researchers stifling. Therefore, design methods for bounded divergence provide practitioners with concepts, frameworks, and tips that structure the scope of solutions while leaving space for creative thinking. In resource-constrained sectors like healthcare and social welfare, this balance is practical: "bluesky" thinking often generates

[8] The cited authors specify that bluesky thinking may be particularly frustrating with marginalized and underserved communities. In this study, we do not make any claims to have worked with any marginalized and underserved groups, nor do we know if any of the participants would consider themselves as such.

solutions that are technically or financially unviable, whereas bounded divergence directs creativity towards interventions that are feasible within the organization's limited capacity.

6.4 Limitations and Future Work

We note three limitations of the present work, regarding the technological scope, domain specificity, and evaluation method, which suggest directions for future research.

First, STCM was designed to support the deployment of discriminative machine learning tools. While interest in generative AI is growing, we scoped our study to discriminative ML because this represented the immediate technological reality of our participant organizations. Consequently, the applicability of the current STCM materials to generative AI remains an open question. However, we anticipate that the underlying theoretical framework remains relevant; challenges such as *Additional Labor* and *Data Inconsistencies* persist regardless of the architecture. Future work should explore how the Sociotechnical Challenge Cards can be refined to explicitly address generative AI, potentially by updating the *Examples* and *Questions* to reflect generative use cases.

Second, we prioritized domain specificity over broad generalizability by grounding STCM in the healthcare and social welfare sectors. We made this choice to ensure the method resonated with the day-to-day realities of our participants. By aligning the content with their specific professional context, we were able to define concrete, recognizable descriptions of challenges and mitigations, avoiding the pitfalls of overly abstract design methods. While this decision limits the immediate transferability of the materials to other sectors (e.g., finance or judicial systems), the structure of the method remains transferable. To facilitate adaptation, we have made the design files for STCM publicly available, enabling researchers and practitioners to tailor the content to new domains.

Finally, constraints on access and availability within these resource-constrained sectors influenced the design of the field experiment. Notably, participants from HC01 were unable to complete the *Countermeasure Planning* activity due to time constraints (limiting the workshop to 90 minutes). Consequently, our findings regarding Objective 2 (Provide Actionable Guidance) are primarily drawn from the ASC01 site. Similarly, limited availability restricted our post-workshop interviews at HC01 to two participants, neither of whom were frontline health professionals. While we mitigated this by triangulating interview data with workshop transcripts and field notes, we acknowledge that the specific perspectives of frontline staff regarding the impact of sociotechnical challenges on patient interaction may be underrepresented. Lastly, due to the project timelines, we were unable to conduct a longitudinal evaluation to assess the long-term impact of STCM. We plan to address these gaps in future work by deploying STCM across a wider range of organizations and conducting longitudinal follow-ups to evaluate the persistence of its outcomes.

7 Conclusion

In this paper, we introduced Sociotechnical Challenge Modeling. STCM is a design method to aid professionals in healthcare and social welfare as they anticipate and address the sociotechnical challenges of machine learning. It consists of a set of Sociotechnical Challenge Cards representing 11 distinct sociotechnical challenges, corresponding Countermeasure Worksheets to support mitigation planning, and Facilitator Guidance and Workshop Templates to enable the adoption of STCM more broadly. Further, we empirically evaluated STCM within two machine learning deployments across England's healthcare and social welfare sectors; these field experiments revealed the perceived value of the method and areas for improvement to be addressed in future work. Our findings, and the implications generated therein, can guide scholars who intend on creating participatory, theory-based design methods.

References

[1] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3290605.3300233

[2] Anita Atwal and Kay Caldwell. 2005. Do all health and social care professionals interact equally: a study of interactions in multidisciplinary teams in the United Kingdom. *Scandinavian Journal of Caring Sciences* 19, 3 (2005), 268–273. doi:10.1111/j.1471-6712.2005.00338.x

[3] Eric P.S. Baumer and M. Six Silberman. 2011. When the implication is not to design (technology). In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2271–2274. doi:10.1145/1978942.1979275

[4] Gordon Baxter and Ian Sommerville. 2011. Socio-technical systems: From design methods to systems engineering. *Interacting with Computers* 23, 1 (2011), 4–17. doi:10.1016/j.intcom.2010.07.003

[5] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John T. Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. *CoRR* abs/1810.01943 (2018). arXiv:1810.01943 http://arxiv.org/abs/1810.01943

[6] Glen Berman, Nitesh Goyal, and Michael Madaio. 2024. A Scoping Study of Evaluation Practices for Responsible AI Tools: Steps Towards Effectiveness Evaluations. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 294, 24 pages. doi:10.1145/3613904.3642398

[7] Claus Bossen, Christian Dindler, and Ole Sejer Iversen. 2016. Evaluation in participatory design: a literature survey. In *Proceedings of the 14th Participatory Design Conference: Full Papers - Volume 1* (Aarhus, Denmark) *(PDC '16)*. Association for Computing Machinery, New York, NY, USA, 151–160. doi:10.1145/2940299.2940303

[8] Robert P Bostrom and J Stephen Heinen. 1977. MIS Problems and Failures: A Socio-Technical Perspective. Part I: The Causes. *MIS Quarterly* 1, 3 (1977), 17–32.

[9] Karen L. Boyd. 2021. Datasheets for Datasets Help ML Engineers Notice and Understand Ethical Issues in Training Data. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 438 (2021), 27 pages. doi:10.1145/3479582

[10] Jonathan Bright, Bharath Ganesh, Cathrine Seidelin, and Thomas M. Vogl. 2019. *Data Science for Local Government*. Technical Report. Oxford Internet Institute.

[11] Bill Buxton. 2010. *Sketching user experiences: Getting the design right and the right design*. Morgan Kaufmann.

[12] Ana Caraban, Loukas Konstantinou, and Evangelos Karapanos. 2020. The Nudge Deck: A Design Support Tool for Technology-Mediated Nudging. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. Association for Computing Machinery, New York, NY, USA, 395–406. doi:10.1145/3357236.3395485

[13] Stevie Chancellor. 2023. Toward Practices for Human-Centered Machine Learning. *Commun. ACM* 66, 3 (2023), 78–85. doi:10.1145/3530987

[14] Hao-Fei Cheng, Logan Stapleton, Anna Kawakami, Venkatesh Sivaraman, Yanghuidi Cheng, Diana Qing, Adam Perer, Kenneth Holstein, Zhiwei Steven Wu, and Haiyi Zhu. 2022. How Child Welfare Workers Reduce Racial Disparities in Algorithmic Decisions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, Article 162, 22 pages. doi:10.1145/3491102.3501831

[15] 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* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 390, 17 pages. doi:10.1145/3411764.3445308

[16] Lucas Colusso, Cynthia L. Bennett, Gary Hsieh, and Sean A. Munson. 2017. Translational Resources: Reducing the Gap Between Academic Research and HCI Practice. In *Proceedings of the 2017 Conference on Designing Interactive Systems*. Association for Computing Machinery, New York, NY, USA, 957–968. doi:10.1145/3064663.3064667

[17] Amanda Coston, Anna Kawakami, Haiyi Zhu, Ken Holstein, and Hoda Heidari. 2023. A Validity Perspective on Evaluating the Justified Use of Data-driven Decision-making Algorithms. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. 690–704.

[18] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. 2020. A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3313831.3376638

[19] 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. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 473–484. doi:10.1145/3531146.3533113

[20] Ying Deng, Alissa N. Antle, and Carman Neustaedter. 2014. Tango Cards: A Card-Based Design Tool for Informing the Design of Tangible Learning Games. In *DIS '14: Designing Interactive Systems Conference 2014*.

[21] Nick Douglas, Nicholas Mays, Mustafa Al-Haboubi, Tommaso Manacorda, Lavanya Thana, Gerald Wistow, and Mary Alison Durand. 2022. Observations of community-based multidisciplinary team meetings in health and social care for older people with long term conditions in England. *BMC Health Services Research* 22, 1 (2022), 758.

[22] Madeleine Clare Elish and Elizabeth Anne Watkins. 2022. *Repairing Innovation: A Study of Integrating AI in Clinical Care.* Technical Report. Data & Society.

[23] Collin J Engstrom, Sabrina Adelaine, Frank Liao, Gwen Costa Jacobsohn, and Brian W Patterson. 2022. Operationalizing a real-time scoring model to predict fall risk among older adults in the emergency department. *Frontiers in Digital Health* 4 (2022), 958663.

[24] Batya Friedman and David Hendry. 2012. The envisioning cards: a toolkit for catalyzing humanistic and technical imaginations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 1145–1148. doi:10.1145/2207676.2208562

[25] Nicola K Gale, Gemma Heath, Elaine Cameron, Sabina Rashid, and Sabi Redwood. 2013. Using the framework method for the analysis of qualitative data in multi-disciplinary health research. *BMC Medical Research Methodology* 13, 117 (2013), 1–8.

[26] Robert D. Galliers and Frank F. Land. 2002. Choosing Appropriate Information Systems Research Methodologies. *Commun. ACM* (2002), 900–902.

[27] 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.

[28] Google. n.d.. Crazy 8's. https://designsprintkit.withgoogle.com/methodology/phase3-sketch/crazy-8s

[29] Christina Harrington, Sheena Erete, and Anne Marie Piper. 2019. Deconstructing Community-Based Collaborative Design: Towards More Equitable Participatory Design Engagements. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 216 (2019), 25 pages. doi:10.1145/3359318

[30] Anna Hedström, Leander Weber, Daniel Krakowczyk, Dilyara Bareeva, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina M.-C. Höhne. 2023. Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond. *Journal of Machine Learning Research* 24, 34 (2023), 1–11. http://jmlr.org/papers/v24/22-0142.html

[31] Alan Hevner and Samir Chatterjee. 2010. *Design Research in Information Systems.* Springer. doi:10.1007/978-1-4419-5653-8

[32] 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*. Association for Computing Machinery, New York, NY, USA, Article 70, 12 pages. doi:10.1145/3419249.3420149

[33] Matthew K. Hong, Adam Fourney, Derek DeBellis, and Saleema Amershi. 2021. Planning for Natural Language Failures with the AI Playbook. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* Association for Computing Machinery, New York, NY, USA. doi:10.1145/3411764.3445735

[34] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3313831.3376219

[35] Anna Kawakami, Amanda Coston, Hoda Heidari, Kenneth Holstein, and Haiyi Zhu. 2024. Studying Up Public Sector AI: How Networks of Power Relations Shape Agency Decisions Around AI Design and Use. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2 (2024). doi:10.1145/3686989

[36] Anna Kawakami, Amanda Coston, Haiyi Zhu, Hoda Heidari, and Kenneth Holstein. 2024. The Situate AI Guidebook: Co-Designing a Toolkit to Support Multi-Stakeholder Early-stage Deliberations Around Public Sector AI Proposals. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3613904.3642849

[37] 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*. Association for Computing Machinery, New York, NY, USA, Article 52, 18 pages. doi:10.1145/3491102.3517439

[38] Finn Kensing and Andreas Munk-Madsen. 1993. PD: structure in the toolbox. *Commun. ACM* 36, 6 (1993), 78–85. doi:10.1145/153571.163278

[39] Miryung Kim, Thomas Zimmermann, Robert DeLine, and Andrew Begel. 2018. Data Scientists in Software Teams: State of the Art and Challenges. *IEEE Transactions on Software Engineering* 44, 11 (2018), 1024–1038.

[40] P. M. Krafft, Meg Young, Michael Katell, Jennifer E. Lee, Shankar Narayan, Micah Epstein, Dharma Dailey, Bernease Herman, Aaron Tam, Vivian Guetler, Corinne Bintz, Daniella Raz, Pa Ousman Jobe, Franziska Putz, Brian Robick, and Bissan Barghouti. 2021. An Action-Oriented AI Policy Toolkit for Technology Audits by Community Advocates and Activists. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 772–781. doi:10.1145/3442188.3445938

[41] Robert E Kraut, Susan R Fussell, Susan E Brennan, and Jane Siegel. 2002. Understanding Effects of Proximity on Collaboration: Implications for Technologies to Support Remote Collaborative Work. In *Distributed Work*. MIT Press, Cambridge, MA.

[42] Sarah Lebovitz, Hila Lifshitz-Assaf, and Natalia Levina. 2022. To Engage or Not to Engage with AI for Critical Judgments: How Professionals Deal with Opacity When Using AI for Medical Diagnosis. *Organization Science* 33, 1 (2022), 126–148. doi:10.1287/orsc.2021.1549

[43] Jane Li, Toni Robertson, Susan Hansen, Tim Mansfield, and Jesper Kjeldskov. 2008. Multidisciplinary medical team meetings: a field study of collaboration in health care. In *Proceedings of the 20th Australasian Conference on Computer-Human Interaction: Designing for Habitus and Habitat* (Cairns, Australia) *(OZCHI '08)*. Association for Computing Machinery, New York, NY, USA, 73–80. doi:10.1145/1517744.1517766

[44] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3313831.3376445

[45] Claire Maiers. 2017. Analytics in action: users and predictive data in the neonatal intensive care unit. *Information, Communication, & Society* 20, 6 (2017), 915–928. doi:10.1080/1369118X.2017.1291701

[46] 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? *Proc. ACM Hum.-Comput. Interact.* 3, GROUP, Article 237 (2019), 23 pages. doi:10.1145/3361118

[47] Tom Mitchell. 1997. *Machine Learning*. McGraw-Hill, New York.

[48] Daniel L Moody. 2003. The Method Evaluation Model: A Theoretical Model for Validating Information Systems Design Methods. In *ECIS 2003 Proceedings*. 17.

[49] Michael Muller. 2008. Participatory Design: The Third Space in HCI. In *The human-computer interaction handbook: fundamentals, evolving technologies, and emerging applications*, Julie A. Jacko and Andrew Sears (Eds.). Lawrence Erlbaum Associates.

[50] Michael J Muller, Jean Hallewell Haslwanter, and Tom Dayton. 1997. Participatory Practices in the Software Lifecycle. In *Handbook of Human-Computer Interaction* (2 ed.), M Helander, T.K. Landauer, and P. Prabhu (Eds.). Elsevier Science, Amsterdam.

[51] Camille Nadal, Shane McCully, Kevin Doherty, Corina Sas, and Gavin Doherty. 2022. The TAC Toolkit: Supporting Design for User Acceptance of Health Technologies from a Macro-Temporal Perspective. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, Article 233, 18 pages. doi:10.1145/3491102.3502039

[52] NHS England. 2024. Artificial Intelligence in Health and Care Award. https://www.england.nhs.uk/aac/what-we-do/how-can-the-aac-help-me/ai-award/

[53] Novia Nurain, Chia-Fang Chung, Clara Caldeira, and Kay Connelly. 2024. Designing a Card-Based Design Tool to Bridge Academic Research & Design Practice For Societal Resilience. In *Proceedings of the 2024 CHI Conference on Human*

Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, Article 817, 20 pages. doi:10.1145/3613904.3642686

[54] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.

[55] Wanda J. Orlikowski. 1992. Learning from Notes: organizational issues in groupware implementation. In *Proceedings of the 1992 ACM Conference on Computer-Supported Cooperative Work*. Association for Computing Machinery, New York, NY, USA, 362–369. doi:10.1145/143457.143549

[56] Wanda J. Orlikowski. 1992. The Duality of Technology: Rethinking the Concept of Technology in Organizations. *Organization Science* 3, 3 (1992), 398–427.

[57] Wanda J. Orlikowski. 2000. Using Technology and Constituting Structures: A Practice Lens for Studying Technology in Organizations. *Organization Science* 11, 4 (2000), 404–428. doi:10.1287/orsc.11.4.404.14600

[58] Abhishek Paul, Roderick Y. Son, Shiv A. Balodi, and Kenney Crooks. 2024. MLOps FMEA: A Proactive & Structured Approach to Mitigate Failures and Ensure Success for Machine Learning Operations. In *2024 Annual Reliability and Maintainability Symposium (RAMS)*. 1–7. doi:10.1109/RAMS51492.2024.10457600

[59] Dorian Peters, Lian Loke, and Naseem Ahmadpour. 2021. Toolkits, cards and games–a review of analogue tools for collaborative ideation. *CoDesign* 17, 4 (2021), 410–434.

[60] 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*. Association for Computing Machinery, New York, NY, USA, 959–972. doi:10.1145/3531146.3533158

[61] Tyler Reinmund, Lars Kunze, and Marina Jirotka. 2024. Transitioning Towards a Proactive Practice: A Longitudinal Field Study on the Implementation of a ML System in Adult Social Care. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. doi:10.1145/3613904.3642247

[62] Tyler Reinmund, Lars Kunze, and Marina Jirotka. 2026. Sociotechnical Challenges of Machine Learning in Healthcare and Social Welfare. arXiv:2601.11417

[63] Tyler John Reinmund. 2023. Designing Data Science Software for Social Care Organisations. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, Article 488, 8 pages. doi:10.1145/3544549.3577062

[64] Shalaleh Rismani, Renee Shelby, Andrew Smart, Edgar Jatho, Joshua Kroll, AJung Moon, and Negar Rostamzadeh. 2023. From Plane Crashes to Algorithmic Harm: Applicability of Safety Engineering Frameworks for Responsible ML. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, Article 2, 18 pages. doi:10.1145/3544548.3581407

[65] Elizabeth B.-N. Sanders, Eva Brandt, and Thomas Binder. 2010. A framework for organizing the tools and techniques of participatory design. In *Proceedings of the 11th Biennial Participatory Design Conference on - PDC '10*. ACM Press, 195–198. doi:10.1145/1900441.1900476

[66] Elizabeth B.-N. Sanders and Pieter Jan Stappers. 2008. Co-creation and the new landscapes of design. *CoDesign* 4, 1 (2008), 5–18. doi:10.1080/15710880701875068

[67] Elizabeth B-N Sanders and Pieter Jan Stappers. 2014. Probes, toolkits and prototypes: three approaches to making in codesigning. *CoDesign* 10, 1 (2014), 5–14.

[68] Devansh Saxena, Ji-Youn Jung, Jodi Forlizzi, Kenneth Holstein, and John Zimmerman. 2025. AI Mismatches: Identifying Potential Algorithmic Harms Before AI Development. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 8, 23 pages. doi:10.1145/3706598.3714098

[69] Mark P Sendak, William Ratliff, Dina Sarro, Elizabeth Alderton, Joseph Futoma, Michael Gao, Marshall Nichols, Mike Revoir, Faraz Yashar, Corinne Miller, et al. 2020. Real-World Integration of a Sepsis Deep Learning Technology Into Routine Clinical Care: Implementation Study. *JMIR Medical Informatics* 8, 7 (2020), e15182.

[70] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, et al. 2023. Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 723–741.

[71] Hong Shen, Wesley H. Deng, Aditi Chattopadhyay, Zhiwei Steven Wu, Xu Wang, and Haiyi Zhu. 2021. Value Cards: An Educational Toolkit for Teaching Social Impacts of Machine Learning through Deliberation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 850–861. doi:10.1145/3442188.3445971

[72] Mona Sloane and Janina Zakrzewski. 2022. German AI Start-Ups and "AI Ethics": Using A Social Practice Lens for Assessing and Implementing Socio-Technical Innovation. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, New York, NY, USA, 935–947. doi:10.1145/3531146.3533156

[73] Alistair Sutcliffe. 2000. On the effective use and reuse of HCI knowledge. *ACM Transactions on Computer-Human Interaction* 7, 2 (2000), 197–221.

[74] Ningjing Tang, Jiayin Zhi, Tzu-Sheng Kuo, Calla Kainaroi, Jeremy J. Northup, Kenneth Holstein, Haiyi Zhu, Hoda Heidari, and Hong Shen. 2024. AI Failure Cards: Understanding and Supporting Grassroots Efforts to Mitigate AI Failures in Homeless Services. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, New York, NY, USA, 713–732. doi:10.1145/3630106.3658935

[75] U.S. Food & Drug. 2024. Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices

[76] John Venable, Jan Pries-Heje, and Richard Baskerville. 2016. FEDS: a Framework for Evaluation in Design Science Research. *European Journal of Information Systems* 25 (2016), 77–89. doi:10.1057/ejis.2014.36

[77] Annika Waern, Paulina Rajkowska, Karin B. Johansson, Jon Back, Jocelyn Spence, and Anders Sundnes Løvlie. 2020. Sensitizing Scenarios: Sensitizing Designer Teams to Theory. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3313831.3376620

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

[79] 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*. Association for Computing Machinery, New York, NY, USA, Article 249, 16 pages. doi:10.1145/3544548.3581278

[80] Richmond Y. Wong, Michael A. Madaio, and Nick Merrill. 2023. Seeing Like a Toolkit: How Toolkits Envision the Work of AI Ethics. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 145 (April 2023), 27 pages. doi:10.1145/3579621

[81] 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*. Association for Computing Machinery, New York, NY, USA, 585–596. doi:10.1145/3196709.3196730

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

[83] Nur Yildirim, Alex Kass, Teresa Tung, Connor Upton, Donnacha Costello, Robert Giusti, Sinem Lacin, Sara Lovic, James M O'Neill, Rudi O'Reilly Meehan, Eoin Ó Loideáin, Azzurra Pini, Medb Corcoran, Jeremiah Hayes, Diarmuid J Cahalane, Gaurav Shivhare, Luigi Castoro, Giovanni Caruso, Changhoon Oh, James McCann, Jodi Forlizzi, and John Zimmerman. 2022. How Experienced Designers of Enterprise Applications Engage AI as a Design Material. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, Article 483, 13 pages. doi:10.1145/3491102.3517491

[84] Nur Yildirim, Changhoon Oh, Deniz Sayar, Kayla Brand, Supritha Challa, Violet Turri, Nina Crosby Walton, Anna Elise Wong, Jodi Forlizzi, James McCann, and John Zimmerman. 2023. Creating Design Resources to Scaffold the Ideation of AI Concepts. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (Pittsburgh, PA, USA) *(DIS '23)*. Association for Computing Machinery, New York, NY, USA, 2326–2346. doi:10.1145/3563657.3596058

[85] 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*. Association for Computing Machinery, New York, NY, USA, Article 356, 13 pages. doi:10.1145/3544548.3580900

[86] 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 Trans. Comput.-Hum. Interact.* 30, 2 (2023). doi:10.1145/3582430

[87] Sabah Zdanowska and Alex S Taylor. 2022. A study of UX practitioners roles in designing real-world, enterprise ML systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, Article 531, 15 pages. doi:10.1145/3491102.3517607

[88] 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. doi:10.1145/1240624.1240704

[89] John Zimmerman, Erik Stolterman, and Jodi Forlizzi. 2010. An Analysis and Critique of Research through Design: towards a formalization of a research

approach. In *Proceedings of the 8th ACM Conference on Designing Interactive Systems*. 310—-319. doi:10.1145/1858171.1858228

A STCM Process

STCM is delivered as a workshop. The workshop is composed of two interconnected activities: Challenge Prioritization and Countermeasure Planning. In Challenge Prioritization, an interdisciplinary group of practitioners – data scientists, managers, senior leaders, frontline care professionals, and end-users – collectively reflect on how the various sociotechnical challenges could affect their ML tool. During Countermeasure Planning, the same group generates unique ideas for addressing high-priority challenges and outlines a mitigation plan. Depending on the time and resource constraints of participating teams, both activities can be held on a single day or spread across multiple. Overall, Challenge Prioritization and Countermeasure Planning require at least 4 hours of dedicated time. STCM can be employed as either a one-off activity or iteratively through several stages of an ML project.

This section defines the process for employing STCM and can serve as instructions for those who intend on employing STCM in their own work.

A.1 Preparation: On-boarding

Before conducting an STCM workshop, there are several steps required to on-board participants and prepare for the workshop activities. First, one stakeholder is identified to act as the facilitator: an individual responsible for delivering the STCM workshop. This person will recruit and on-board participants, schedule the workshop activities, facilitate the workshop, and document and disseminate the workshop outputs. Second, the facilitator recruits participants for the workshop. STCM is an interdisciplinary activity, requiring the expertise and experiences of a wide range of stakeholders: frontline care professionals, data scientists, engineers, project managers, and senior leadership. While recruiting participants, the facilitator – with the support of senior leaders and key stakeholders – identifies a time and place to conduct the workshop; during scheduling, the facilitator should consider whether it is feasible to conduct the workshop within a single day or if it is necessary for the workshop to be held across multiple.

After scheduling the workshop, the facilitator provides instructions to the participants on the objectives, activities, and outcomes of STCM. First, the facilitator sends instruction materials to the participants. The instruction materials provide guidance on how to use STCM, specific guidance on the workshop activities, and an introduction to key concepts: the ML-enabled Care Pathway and the 11 sociotechnical challenges. The instruction materials are available on the STCM website where they can also be downloaded as a PDF. Additionally, at the beginning of the workshop, the facilitator conducts an introductory presentation to review the same content and address participant questions. The introductory presentation is also available on the STCM website.

A.2 Activity 1: Challenge Prioritization

After on-boarding, participants engage in the first activity: Challenge Prioritization. In this activity, participants discuss the sociotechnical challenges that could arise on their ML project and prioritize them along a customizable likelihood by severity scale. First, the facilitator leads the participants in constructing a likelihood by severity scale on a sheet of paper; this scale is used to rank the sociotechnical challenges. The concept of a likelihood by severity scale is informed by risk management methods such as Failure Modes and Effects Analysis (FMEA), a method that has been found to have potential use in social and ethical risk assessments for ML [58, 64]. There are two reasons for the use of a prioritization scale. First, it is unlikely that all sociotechnical challenges are relevant to every ML tool. For example, Poor Environmental Conditions are likely to only impact ML tools which consume data that are susceptible to variations in environmental factors. Similarly, teams may already have mitigations in place to address some of the challenges and therefore require less attention. Second, healthcare and social welfare organizations are resource- and time-constrained; attempting to address each challenge would prove infeasible and unnecessarily onerous. As organizations may employ idiosyncratic prioritization scales within existing processes, STCM does not prescribe levels for likelihood or severity but instead relies on participants to define them in accordance with internal standards.

After constructing the prioritization scale, the facilitator divides participants into groups of 2 or 3. It is important to ensure that the subgroups are evenly distributed in terms of experience with the ML tool and discipline. This allows participants to approach the sociotechnical challenges from a diversity of perspectives. For instance, those with limited experience or background knowledge of the ML tool should not be paired together; similarly, data scientists should not be paired with data scientists, and GPs with GPs. The facilitator distributes the Sociotechnical Challenge Cards to each subgroup and selects an initial challenge for discussion. In their subgroups, participants read the card independently before discussing with the other members of the subgroup. During the subgroup discussions, the facilitator instructs the participants to consider the following prompts: *What would cause this challenge to arise? How likely is this challenge to arise? If this challenge were to occur, what would be its consequences? How severe would those consequences be?*

Following the subgroup discussions, the facilitator initiates a discussion between the broader group, eliciting reflections from several participants. At this stage, participants rank the challenge according to the likelihood of it arising and the severity of its consequences by placing the cards along the prioritization scale. The facilitator elicits rationales for participants' decisions.

Once a challenge is placed on the scale, the facilitator selects the next challenge and repeats the process until all 11 challenges have been discussed. To conclude Activity 1, participants select 3 to 4 high-priority challenges to discuss in the second activity.

A.3 Activity 2: Countermeasure Planning

The second activity – Countermeasure Planning – aims to create an actionable plan to address the prioritized sociotechnical challenges. Once again, the facilitator divides participants into subgroups: one subgroup for each prioritized challenge. In their subgroups, participants first review the sociotechnical challenge to develop a shared understanding and resolve any uncertainties.

After an initial discussion, the facilitator provides each participant with a Countermeasure Worksheet. Participants perform two steps with the aid of the worksheet. First, participants perform the adapted Crazy 8's countermeasure ideation technique. In this step, participants generate eight distinct countermeasures for their respective sociotechnical challenge, spending 1 minute on each idea for a total of 8 minutes. Whereas conventional Crazy 8's methods prompt participants to sketch each idea, this exercise only requires participants to write them down. This is done to decrease the barrier to participation for those who have limited sketching experience. Once participants have completed this step, they present their ideas to the other members of their subgroup. Collectively, participants select at least three of the most compelling ideas, according to (i) what is likely to have the greatest impact on the challenge and (ii) what is most feasible to implement. Next, participants review the predefined countermeasures available on the backside of the Countermeasure Worksheet. Participants revise the list of countermeasures where appropriate, adding those that are relevant or integrating different countermeasures together.

To conclude the activity, each of the subgroups presents their proposed countermeasures to the others, soliciting feedback and questions. The facilitator documents the outcomes of this discussion with the aid of the templates available on the STCM website.

B Sociotechnical Challenge Cards

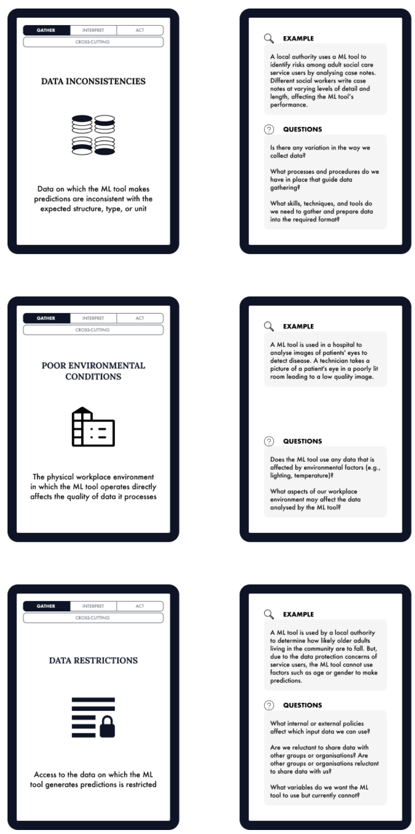


Figure 4: Sociotechnical Challenge Cards that arise during the "Gather" stage.

Figure 5: Sociotechnical Challenge Cards that arise during the "Interpret" stage.

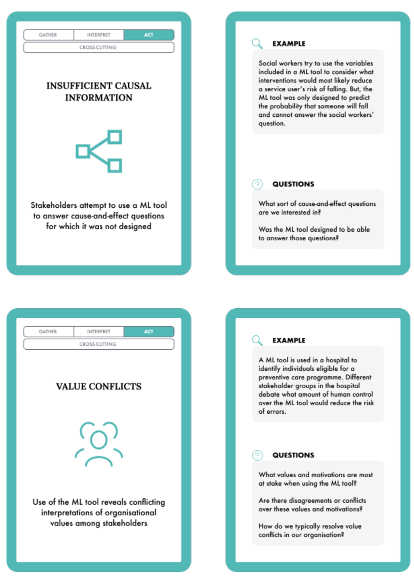


Figure 6: Sociotechnical Challenge Cards that arise during the "Act" stage.

Figure 7: Sociotechnical Challenge Cards that arise across multiple stages.