Empowering Stakeholders with Participatory Auditing of Predictive AI: Perspectives from End-Users and Decision Subjects without AI Expertise

Patrizia Di Campli San Vito*
School of Computing Science
University of Glasgow
Glasgow, United Kingdom
patrizia.dicamplisanvito@glasgow.ac.uk

Eva Fringi*
School of Computing Science
University of Glasgow
Glasgow, United Kingdom
evangelia.fringi@glasgow.ac.uk

Penny S. Johnston
Computing Science and Mathematics
University of Stirling
Stirling, United Kingdom
Penny.Johnston@stir.ac.uk

Leonardo C. T. Bezerra
Computing Science and Mathematics
University of Stirling
Stirling, United Kingdom
leonardo.bezerra@stir.ac.uk

Marios Aristodemou
University of York
York, United Kingdom
marios.aristodemou@york.ac.uk

Siamak F. Shahandashti
Department of Computer Science
University of York
York, United Kingdom
siamak.shahandashti@york.ac.uk

Emily O'Hara
University of Sheffield
Sheffield, United Kingdom
e.w.ohara@sheffield.ac.uk

Laura Fiona Whyte
School of Computing Science
University of Glasgow
Glasgow, United Kingdom
laura.f.whyte@glasgow.ac.uk

Lin Luo
School of Computing Science
University of Glasgow
Glasgow, United Kingdom
2841697L@student.gla.ac.uk

Mark Wong
Division of Urban Studies and Social Policy
University of Glasgow
Glasgow, United Kingdom
Mark.Wong@glasgow.ac.uk

Ayah Soufan
University of Strathclyde
Glasgow, Scotland, United Kingdom
ayah.soufan@strath.ac.uk

Yashar Moshfeghi
NeuraSearch Lab
University of Strathclyde
Glasgow, United Kingdom
yashar.moshfeghi@strath.ac.uk

Simone Stumpf
School of Computing Science
University of Glasgow
Glasgow, United Kingdom
simone.stumpf@glasgow.ac.uk

Figure 1: Current AI auditing involves typically only technical experts with AI background. The novel concept of Participatory AI auditing involves end-users and decision subjects who typically have no AI background.

*Shared first authors

CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3791757

Abstract

Artificial intelligence (AI) applications have become ubiquitous in their impact on individuals and society, highlighting a crucial need for their responsible development. Recent research has called for *participatory AI auditing*, empowering individuals without AI expertise to audit AI applications throughout the entire AI development pipeline. Our work focuses on investigating how to support these kinds of auditors through participatory AI auditing tools and processes. We conducted a series of co-design workshops, using two health-related predictive AI applications as examples. Our results show that participants wanted to be part of AI audits, and were insightful in identifying the potential impacts of applications, but needed to be assisted in conducting audits, especially how to measure impacts. Importantly, participants provided examples of impacts not considered in current risk/harm taxonomies. Our findings provide implications for the design of tools and processes to empower everyone to contribute to responsible AI development in the future.

CCS Concepts

• **Human-centered computing** → **User studies**; • **Computing methodologies** → **Artificial intelligence**.

Keywords

Predictive AI, Participatory Auditing, Co-Design, Responsible AI, Harms, Benefits, Health, End-Users, Decision Subjects

ACM Reference Format:

Patrizia Di Campli San Vito, Eva Fringi, Penny S. Johnston, Leonardo C. T. Bezerra, Marios Aristodemou, Siamak F. Shahandashti, Emily O'Hara, Laura Fiona Whyte, Lin Luo, Mark Wong, Ayah Soufan, Yashar Moshfeghi, and Simone Stumpf. 2026. Empowering Stakeholders with Participatory Auditing of Predictive AI: Perspectives from End-Users and Decision Subjects without AI Expertise. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 35 pages. https://doi.org/10.1145/3772318.3791757

1 Introduction

Predictive Artificial Intelligence (AI) applications[1] are having a significant impact on individuals and society, for example, in loan applications [50] and healthcare [54]. Spectacular failures, such as blatant gender bias in AI used to make hiring decisions [48] and racism perpetuated in bail decisions [43], have underscored the need for responsible AI development [1, 19, 33]. Alongside these research efforts, emerging regulation, such as the EU AI Act [49] and the international Hiroshima Policy Framework [29], have cemented the requirement for responsible approaches to AI. To mitigate the risks and harms posed by AI applications, regular assessments throughout the AI development pipeline are necessary to ensure accountability.

In this context, AI auditing [6] is a new area addressing accountability by providing a "*systematic, independent and documented process for obtaining audit evidence and evaluating it objectively to determine the extent to which the audit criteria are fulfilled*" [34]. Although the specifics of the AI audit process can differ significantly from traditional auditing, the overall AI audit structure tends to follow a common set of stages: *Harms Discovery*, *Standards Identification*, *Performance Analysis*, and *Audit Communication and Advocacy* [6].

AI audits are typically conducted by individuals with an AI background, whom we refer to as AI experts, often within regulatory bodies or by AI developers themselves. However, in a shift towards more responsible, human-centred AI development practices [5], there have been efforts to introduce *participatory AI auditing* [5, 23, 24, 38], in which stakeholders who typically do not have AI expertise, such as end-users, who use the AI application, and decision subjects, who the AI application is making predictions about, are empowered to take the role of an auditor, see Figure 1. This follows research showing that these participatory auditors have the potential to uncover issues often overlooked by AI experts [3, 16, 22, 23, 38, 40, 55]. However, there is currently a lack of auditing tools that are suitable for end-users and decision subjects, whereas AI experts are supported by a growing list of tools to assist with (often fairness-related) aspects of the assessment of applications [13, 15, 31, 42, 46, 61, 63]. While a growing body of work investigates the involvement of end-users and decision subjects in responsible AI development [23, 37, 45], there is limited research to directly involve them through participatory AI auditing tools.

In this paper, we address this challenge: *How can we directly involve individuals without AI expertise in the whole AI auditing process? How can we support auditors at each stage of a participatory audit through tools?* To investigate this, we ran a series of nine co-design workshops to collect feedback from potential participatory auditors. We grounded our research in two health-related predictive AI applications.

Our results show that there is a clear role for end-users and decision subjects as part of a group of external auditors, moving away from current auditing practices relying solely on AI experts. Participants highlighted the need for participatory auditing throughout the AI development pipeline, not just at the end or after deployment. Our participants requested a wide range of information to understand AI applications and their impact, which has to be provided in a way which allows individuals without an AI background to understand. Participatory auditors need to be supported systematically through an audit and want the ability to not only include negative, but also positive implications of an AI application. While taxonomies can be useful, they should not be restrictive and instead be used as thinking aids. Participatory auditors need support through a tool especially when creating metrics to measure the implications they foresee for an AI application.

This paper makes the following contributions, which will benefit researchers and practitioners in AI accountability and responsible AI development as well as designers and developers of tools to support participatory auditing, by:

- Demonstrating the value of participatory auditing from the point of view of end-users and decision subjects;
- Defining approaches for the participatory AI auditing process;
- Providing a better understanding of the information and support needs of participatory auditors during auditing;

[1]We distinguish predictive, often also called discriminative AI, from generative AI. We only deal with the former in this paper.

- Shaping directions for the design of participatory auditing tools.

2 Related Work

Our research is building on existing work in 1) AI auditing, including harm and impact identification, 2) the novel perspective of participatory auditing, and 3) emerging tools and user interfaces that support the auditing of AI.

2.1 AI Auditing

Auditing is defined as a "*systematic, independent and documented process for obtaining audit evidence and evaluating it objectively to determine the extent to which the audit criteria are fulfilled*" [34]. Audits can either be internal - conducted by or on behalf of the organisation itself - or external - conducted either by a party with an interest in the organisation or an independent auditing organisation.

Though auditing is a long-established field, AI auditing is a much more recent topic that is attracting significant attention as responsible AI efforts increase, partly driven by emerging regulation addressing accountability, such as the EU AI Act [49]. Birhane et al. [6] conducted a systematic review of over 300 works documenting AI auditing practices carried out by both academic and non-academic organisations. Their results show that AI can be categorised as a function of the audit scope: (i) model, which may be broken down into data and/or algorithm; (ii) product, and; (iii) ecosystem, which in addition to the product also includes communities and sociotechnical environments. The review showed that AI audits typically follow a four-stage process involving Harms Discovery, Standards Identification, Performance Analysis, and Audit Communication and Advocacy [6].

Recent research has focused extensively on harm identification and impact assessments as part of AI auditing. An influential domain taxonomy was proposed by Weidinger et al. [62]; this taxonomy comprises six areas of harms: (i) representation & toxicity; (ii) misinformation; (iii) information & safety; (iv) malicious use; (v) human autonomy & integrity, and; (vi) socioeconomic & environmental. The MIT AI Risk Repository [56] was built on this earlier taxonomy and accommodates over 1600 risks observed across 65 taxonomies.

2.2 Participatory AI Auditing

Underlying the move towards participatory auditing, there have been recent calls to innovate AI development practices, which currently are "*technically focused, representationally imbalanced, and non-participatory*" [5], promoting a paradigm shift towards end-users and decision subjects. This has resulted in the novel concept of *participatory auditing* [5, 23, 24] in which end-users and decision subjects who do not have AI expertise are empowered to take the role of an auditor.

Recent research on AI auditing has established the effectiveness of end-user and decision subjects' contributions in detecting and analysing harmful AI behaviours [38]. In many cases, individuals without formal training in AI have been successful at identifying potentially harmful AI biases, which had previously escaped the experts' attention [23, 55]. As a result, AI experts are increasingly motivated to engage end-users and decision subjects in AI auditing to help overcome any blind-spots persisting after technical approaches [22]. In fact, Lam et al. [39] argue that end-users and decision subjects should be seen as intrinsic to AI audits as an integrated, and dynamic part of the audit process. Moreover, Tang et al. [58] have demonstrated that enabling end-users and decision subjects to understand AI failures can elicit constructive feedback and useful mitigation strategies. Extending these insights, Solyst et al. [57] provide evidence that, with appropriate scaffolding, adolescents can serve as valuable auditors who often engage with emerging AI technologies more rapidly or differently than adults, thus advancing broader critical computing education efforts in AI literacy. These developments have highlighted the need for robust methodologies and tools designed to support end-users and decision subjects in conducting AI audits.

2.3 Tools for Predictive AI Auditing

Existing predictive AI auditing tools predominantly target AI experts. These tools offer technical metrics and visualisations to diagnose fairness or robustness issues. For example, AI Fairness 360 [4] and Fairlearn [61] provide bias detection, mitigation algorithms, and fairness evaluation capabilities, but require programming expertise. Other tools like the What-if Tool [63], FairHIL [46], and FAIRVIS [15] allow AI experts to explore fairness through counterfactual reasoning, subgroup analysis, and causal graph inspection. Finally, tools such as Silva [64] focus on uncovering bias in model reasoning.

Research on auditing frameworks that target individuals without AI expertise is starting to emerge but is mainly focused on addressing fairness. One example is the Stakeholders' Fairness Framework [18], which combines interactive interfaces with interview protocols to elicit fairness perspectives. Another example is IndieLabel [38], which enables marginalised communities without coding expertise to audit toxicity models. Similarly, FairHIL [47] supports fairness assessments and allows user feedback.

More recently, emphasis has been given to developing tools that involve end-users and decision subjects in impact assessments [8, 14]. These tools use a minimal amount of stakeholder input, which combined with large language models (LLMs) automatically generates impact assessments for AI applications, aiming to reflect stakeholders' ethical concerns, perceived risks, and benefits. They additionally include a phase in which stakeholders review the AI-generated content. Nevertheless, such approaches may be overly streamlined, limiting opportunities for auditors to think creatively or contribute original insights.

There is currently a severe lack of tools for participatory auditing that support a structured process for conducting comprehensive audits. Designing such tools could support a more holistic approach to predictive AI auditing that engages diverse stakeholders and connects technical assessments to broader understandings of impact and accountability.

3 Methods

To answer our research questions, we employed a co-design approach [9, 52] in which we collaborated with 17 participants without

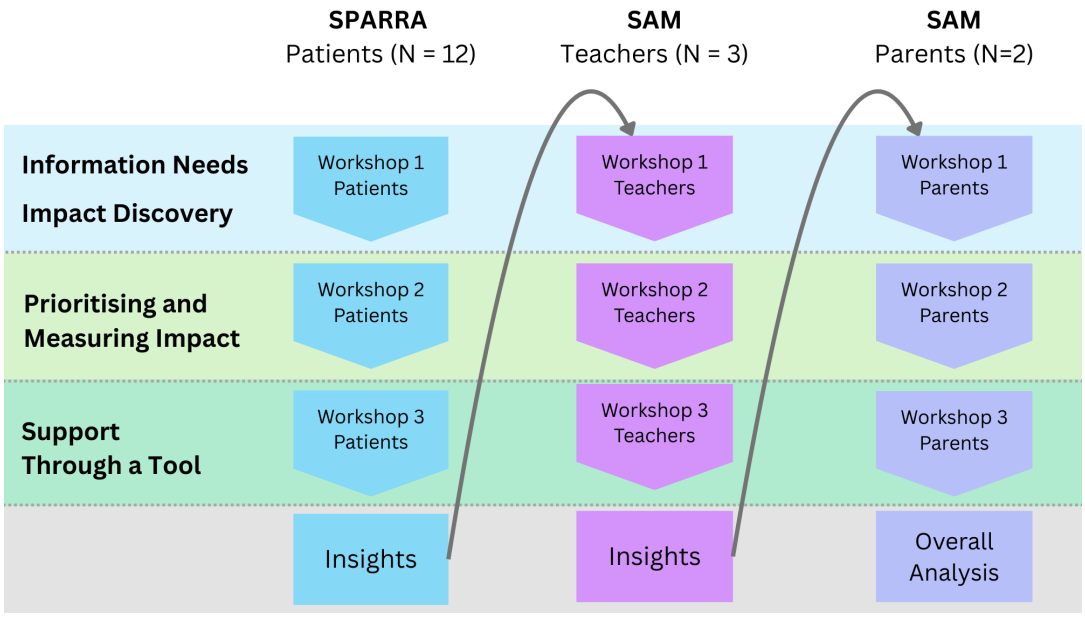


Figure 2: Overview of our co-design workshop structure. We conducted nine co-design workshops with various target stakeholder groups. We drew on insights from earlier workshops to inform the topics of subsequent ones.

AI expertise, representing potential end-users and decision subjects, in a series of nine workshops. We grounded our investigations in two real AI applications: Application 1: *Scottish Patients at Risk of Re-Admission and Admission (SPARRA)* , which predicts a score about a patient's risk of being admitted to hospital based on their health and demographic records, and Application 2: *School Attachment Monitor (SAM)* , which predicts secure or insecure attachment styles from speech recordings of children. The overall structure of the workshops is given in Figure 2. We conducted three workshops around *SPARRA* with 12 patient representatives (Workshops labelled Patients). We then worked with three teachers (Workshops labelled Teachers) and two parents of children (Workshops labelled Parents) on *SAM*, in three workshops for each group.

Across the nine workshops, we allowed insights from earlier workshops to shape the content of later ones. Importantly, while there were shared topics for discussion and shared activities across workshops (conducted between different participant groups), they were adapted based on what we learnt previously. The user research was approved by the University of Glasgow, College of Science and Engineering Ethics Committee. In this section, we introduce the AI applications to ground our research. Then, we detail workshop participants and structure. Finally, we briefly discuss how we analysed the data obtained from the workshops.

3.1 The AI Applications

3.1.1 AI Application 1: Scottish Patients at Risk of Readmission and Admission (SPARRA). *SPARRA* is a long-standing, deployed predictive AI model designed to help general practitioners (GPs) in Scotland identify patients at high risk of emergency hospital admission within the subsequent 12-month period, enabling targeted interventions and care planning, which may potentially reduce emergency admissions and improve patient outcomes. This model is run monthly, resulting in scores for approximately 4.3 million individuals. The model outputs scores ranging from 1 to 99, with higher values indicating increased likelihood of emergency admission.

SPARRA was constructed using logistic regression on comprehensive patient-level data, including inpatient admissions, prescriptions dispensed in the community, emergency department attendances, new outpatient attendances, and psychiatric inpatient admissions. Training data, access to the model, or outputs were not available; hence, this is an instance of 'opaque-box auditing' [17] in which auditors have limited information to inspect.

3.1.2 AI Application 2: School Attachment Monitor (SAM). *SAM* is a research prototype, developed by the University of Glasgow [12], which comprises a predictive AI model aimed at helping child

psychologists and psychiatrists to screen for cases of insecure attachment in children between 5 and 9 years old, indicating when intervention might be needed. Attachment styles [10] - bonds between infants and their caregivers - are important to later relationships, emotional, physical and mental health, overall well-being and quality of life [25, 28, 41].

The dataset underlying the model consists of 512 recordings of 104 children aged between 5 and 9 years, undergoing the Manchester Child Attachment Story Task (MCAST) [30], which examines how children elaborate on five different hypothetical stressful scenarios as an attachment assessment test. Each recording was assessed independently by two experts and a label - *secure attachment* or *insecure attachment* - was agreed between them. The audio data from the dataset were used to train a multimodal Recurrent Neural Network (RNN), which combines speech transcripts and acoustic features. The data is protected but access to the model is feasible; this means that auditors can conduct a 'translucent-box audit' [17] where they can probe the model.

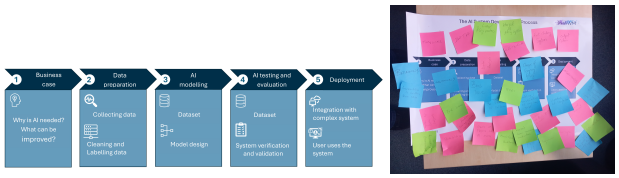
3.2 Setup of the Co-design Workshops

We leveraged the four-stage participatory AI auditing process proposed by Birhane et al. [6] - *Harms Discovery*, *Standards Identification*, *Performance Analysis*, and *Audit Communication and Advocacy* - to set up our workshops. We decided to rephrase the first stage from Harms Discovery to Impact Discovery, which also allows for neutral or positive views of AI applications.

The general topics addressed at the workshops were as follows: Workshop 1 addressed the auditing process in general and the information needs of participatory auditors, as well as an initial Impact Discovery. Workshop 2 focused on how to prioritise and measure impacts, to account for the Standard Identification stage of the process. Finally, Workshop 3 concentrated on tool support for the auditing process, including Performance Analysis and Audit Communication and Advocacy, using 'provotypes' [7] to obtain feedback.

The workshops for *SPARRA* were conducted between October 24 and December 18, 2024; and for *SAM*, between May 14 and June 20, 2025. Each workshop lasted two hours. Table 1 offers a summary of the workshops' design, outlining the co-design activities and protocol changes between workshops.

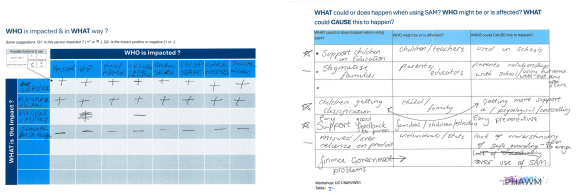
3.2.1 Participants. For *SPARRA*, we recruited 12 participants from 96 applicants, through posters, social media, and mailing lists of local community organisations. All resided in Scotland. Because *SPARRA* is aimed at a broad range of patients, we aimed at assembling a diverse group of participants from diverse ethnic backgrounds, income levels, and ages who could potentially be affected by the application's predictions. Thus, we recruited four cisgender female, one transgender female, six cisgender male and one transgender male participants; aged between 20 and 74 years (average age 34.8 ± 15.3); six reported ethnicities; a range of sexualities; different caring responsibilities; varying familiarity with algorithms and Tech Savviness [53], but little familiarity with AI. Due to the size of the workshops for *SPARRA*, we divided participants into smaller subgroups (varying between workshops) for activities to facilitate discussions. We encountered some drop-out; only 11 participants returned for Workshop 2 and 10 for Workshop 3.



(a) AI Development Pipeline used in Workshop 1, adapted from Vincenzi et al. [59] with permission.

(b) AI Development Pipeline with post-its of what information was needed (pink), when the application should be audited (blue) and who should audit it (green)

Figure 3: AI Development Pipeline without and with post-its of Workshop 1 Patients. See larger versions in Appendix E Figure 11.



(a) Impact Assessment Matrix

(b) Event Table

Figure 4: Impact Assessment Matrix from Workshop 1 Patients and Event Table from Workshop 1 Teachers. See larger versions in Appendix E Figure 12.

The recruitment process for the *SAM* workshops focused on stakeholders with a vested interest in the outcome of a prediction application such as *SAM*, which we identified as parents of children within the target age range and their teachers. We recruited a group of three teachers connected to a local primary school (all white and female, two aged between 25–35 and one aged between 45–54) and a group consisting of two parents (one male and one female, both white and aged between 35-44) through project connections and institution-led online platforms. Both groups had little familiarity with AI but had high technological savviness.

All participants were compensated with £40 per workshop, apart from the three teachers who participated in a professional capacity during contracted hours.

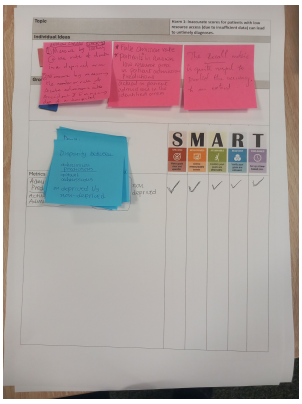
3.2.2 Workshop 1: Participatory auditing process, information needs and initial Impact Discovery. In designing activities, we drew heavily on the AI development pipeline as suggested by Vincenzi et al. [59], see Figure 3(a), as well as the auditing process suggested by Birhane et al. [6].

In **Workshop 1 Patients**, aimed at *SPARRA*, we first introduced auditing in the context of generic predictive AI applications to capture overall thoughts and attitudes on the auditing process for predictive AI applications in general. In Activity 1, participants were then asked to place post-its on the appropriate stage of a poster of the AI development pipeline, identifying (i) what information they

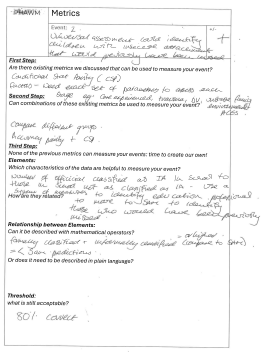
	Activity	Application & Stakeholder Group	Artefacts	Reflection Type	Changes for the following Group
WS1	1. Information Needs • What Info • When & Why • Who	App 1 - Patients	Post-its on pipeline	Individual	• Shifted focus form generic predictive AI to specific AI application
		App 2 - Teachers			
		App 2 - Parents			
	2. Impact Discovery	App 1 - Patients	Impact assessment matrix	Individual & Group	• Harms → Events • Matrix → Event Table
		App 2 - Teachers	Event table		• Introduced Taxonomy
		App 2 - Parents			
WS2	1. Prioritise Events	App 1 - Patients	Menti-meter ratings	Individual	• Impact → magnitude • Included both positive and negative events
		App 2 - Teachers			
		App 2 - Parents			
	2. Create Metrics	App 1 - Patients	Post-its	Individual & Group	• Included both positive and negative events • Different metrics for App2 • Changed SMART framework to 3 step approach
		App 2 - Teachers	Metric forms	Group	• Cut down existing metrics • Presentation in steps was confusing → showed them side by side next
		App 2 - Parents			
WS3	1. Audit tool screen flow and information	App 1 - Patients	Screen Info pages	Group	• Presented user flow from WS1 - Patients
		App 2 - Teachers	Post-its and notes on provotypes		
		App 2 - Parents			
	2. Design audit tool screens	App 1 - Patients	Wireframe sketches	Group	• Created wireframes on the sketches as provotypes
		App 2 - Teachers	Post-its and notes on provotypes		• Discussed suggested changes in the next WS
		App 2 - Parents			

Table 1: Workshop summary describing the activities of each workshop, with which group each activity was conducted, the artefacts resulting from the activity, if the activity was conducted individually or in the group (reflection type), and what changes where implemented between the groups.

would need to audit a predictive AI application, (ii) when it should be audited (and why then), and (iii) who should audit, see Figure 3(b). We then introduced the specifics of *SPARRA*. Participants then completed Activity 2, an impact assessment matrix for *SPARRA*, see Figure 4(a), associating potential impacts, who would be impacted and how, as well as the positive/negative nature of the impact. In the same activity, we asked participants to report any harms or negative outcomes they had observed and collected these in a whole-group discussion during the workshop.



(a) Metric Creation with SMART from Workshop 2 Patients. See larger version in Appendix E Figure 13.



(b) Metric Creation with three cases as steps from Workshop 2 Teachers. See larger version in Appendix E Figure 14.

Figure 5: Metric Creation templates from Workshop 2 Patients and Workshop 2 Teachers.

Insights and Changes: Activity 1 collected a broad range of information about *generic* predictive AI auditing. Therefore, we decided to modify this activity to focus on the specific example of *SAM* during Workshop 1 Teachers. We also changed the terminology used in Activity 2. Rather than focusing on harms, we referred to impacts as *Events*, to accommodate the possibility that they could be either positive (benefits) or negative (harms). We also modified the matrix to an event table, see Figure 4(b), to ease problems with filling out the same information repeatedly.

In **Workshop 1 Teachers**, as previously, we first introduced the participants to AI auditing and predictive AI. Next, we discussed *SAM*. Then, in Activity 1, participants placed post-its directly on the AI development pipeline depicting the details of *SAM*. Afterwards, in Activity 2, participants described events , their affected entities, and their potential causes.

Insights and Changes: The Event table led to more focused and better-defined events. As previous work suggested the usefulness of a taxonomy during the event creation process [37, 60], in Workshop 1 Parents we introduced and discussed the taxonomy proposed by Weidinger et al. [62] to hear participants' feedback. We then asked participants to classify their created events according to this taxonomy at the end of Activity 2.

For **Workshop 1 Parents**, Activities 1 and 2 were conducted in the same way as Workshop 1 Teachers, but then we introduced the taxonomy and asked participants to classify the events they had created using this taxonomy. An additional benefit of using a taxonomy was that participants discovered events they had not previously anticipated. We therefore allowed them to create more events after seeing the taxonomy.

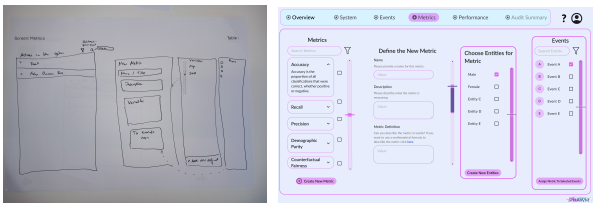
Insights and Changes: The taxonomy supported participants in thinking about a wider range of events and classifying some of their events.

3.2.3 Workshop 2: Prioritising and Measuring Impacts. To address Standards Identification [6], in **Workshop 2 Patients**, researchers returned to the list of the negative events (harms) created in Workshop 1 Patients for *SPARRA*. The list combined similar events and rephrased them for clarity. In Activity 1, participants were asked to rate each event in terms of its likelihood and impact. Then, researchers introduced the idea of metrics and what metrics had already been established for *SPARRA*, discussing in detail three performance metrics (False Omission Rate, Recall and False Discovery Rate) and two fairness metrics (Demographic Parity and False Discovery Rate Parity for gender). Participants were then introduced to the SMART framework [26], in order to prompt them to create **S**pecific, **M**easurable, **A**ttainable, **R**elevant and **T**ime-based metrics. In Activity 2, participants were asked to create metrics for events for the remainder of the workshop, starting with one of the highest priority ones. They did so first individually using post-its, then further consolidating as a group (Figure 5(a)).

Insights and Changes: For prioritisation, we changed 'impact' to magnitude (of impact) for the following workshops to make it more precise. We decided to include both positive and negative events in the prioritisation and metric creation process. Researchers changed metrics for *SAM* because different metrics had been used in *SAM* in the past. We found that creating metrics was challenging for participants, and the SMART framework did not adequately support them in this task. Hence, we changed the metric creation process, checking whether 1) existing metrics could be used to measure an event, 2) combinations of these existing metrics could be used, or 3) whether a new metric needed to be created, following an approach inspired by Nakao et al. [46].

In **Workshop 2 Teachers**, participants prioritised the events from Workshop 1 in Activity 1. Researchers then introduced metrics of *SAM* that had been applied in the past: Accuracy, Recall, and Precision as performance metrics, and Demographic Parity, Conditional Statistical Parity, and Accuracy Parity for gender as fairness metrics. For Activity 2, participants then focused on creating metrics for two events (for example, Figure 5(b)), which had been rated as having high priority. To support Activity 2, researchers asked as a first step, if already discussed metrics could be used to measure the event, or, as a second step, if their combinations could be used. If not, in step three, a new metric needed to be created and participants were led through a process inspired by the metric builder by Nakao et al. [46]. Participants began by identifying relevant characteristics of the AI application's data and then considered the nature of the relationship between these characteristics, distinguishing between quantitative and qualitative measures. Finally, after designing the metric, they were asked to specify acceptable thresholds. This helped participants to define the criteria for what should be considered an acceptable outcome when auditing the event.

Insights and Changes: We decided to cut down existing metrics to four because explaining all six took significant time without providing a clear benefit. We found that the metric creation process worked better, but the presentation in steps was confusing. Therefore, we showed the options alongside each other.



(a) Wireframe sketch from Workshop 3 Patients.

(b) Wireframe provotype from Workshop 3 Teachers/Parents.

Figure 6: Wireframe sketch from Workshop 3 Patients and wireframe provotypes used in Workshop 3 Teachers/Parents, showing the Metrics screen with a list of metrics already in the system and a mechanism to create new metrics. The main elements of the screens are the same, with only small changes between the sketch and the wireframe, such as added checkboxes for the already existing metrics. See larger versions in Appendix E Figure 15.

In **Workshop 2 Parents**, Activity 1 was conducted unchanged to prioritise Events. Then the four metrics (Accuracy, Recall/Accuracy Parity, Demographic Parity) were explained. Participants worked on Activity 2 to create metrics for two events.

Insights and Changes: The presentation of the metric creation steps alongside each other supported participants better, but the metric creation process was still challenging for them.

3.2.4 Workshop 3: Support through a tool. For **Workshop 3 Patients** Activity 1, participants were asked to discuss the order in which they would like to go through the tool, using screen outlines containing only their screen names and sample information from Workshop 1 Patients, following the audit process suggested by Birhane et al. [6] (see details in Appendix A). As they worked through the flow of screen outlines, they were also asked to attend to the information on each screen, changing/adding information or adding additional screens. During Activity 2, participants were asked to design the main screens for the tool, with researchers following participants' instructions to complete the screen outlines (see Figure 6(a)). To support this activity, researchers provided low-fidelity prototyping materials [51], such as common user interface elements (for example, drop-down menus, buttons, and sliders), along with *SPARRA* information and event and metric details identified in the previous workshops, which could be dropped into the screen outlines.

Insights and Changes: Instead of making participants design the flow and screens from scratch, a significant effort for them, we instead mapped the user flow from Workshop 1 Patients and presented that flow, as well as created wireframes on the sketches as provotypes [7] for the following workshops.

In **Workshop 3 Teachers**, participants first discussed the user flow as Activity 1, providing feedback, changes and clarification. They also returned to information they requested in Workshop 1, to better understand what information they deemed necessary to audit and where it would fit in the process. In Activity 2, participants reviewed the wireframes (see Figure 6(b)), providing detailed feedback and suggesting changes.

Insights and Changes: Overall, the previous changes we made were well-received. We did not update the wireframes for the last workshop, however, we discussed suggested changes with participants.

In **Workshop 3 Parents** Activity 1, participants discussed the user flow, screen names and suggested changes. As in the previous workshop, they also assigned the information they requested in Workshop 1 to the screens. In Activity 2, they discussed and fed back on the wireframes and proposed changes.

Insights and Changes: Parents proposed changes and agreed with some changes proposed by the Teachers, but not all.

3.3 Data Analysis

During all workshops, we collected artefacts during co-design activities, such as post-its and participant writings, and recorded all discussions to supplement the artefacts. We then conducted a reflexive thematic analysis [11] on the artefacts, supplemented by transcripts of the recordings: for Workshop 1, we analysed the AI development pipeline posters with the three-coloured post-its, the consolidated lists of events and the classification to the taxonomy by Weidinger et al. [62], while for Workshop 2, the metric creation themes and strategies (for Workshop 2 Patients, where the process was not as guided) were analysed. The two lead authors coded the artefacts individually and then discussed the themes until a consensus was reached. Each post-it was counted as new input and coded. Similarly, themes were coded and their occurrences were counted based on the number of entries in the rest of the artefacts (presented with *N=x*). The event prioritisation of Workshop 2, the information participants wanted to see on screens and the sketched screens, as well as the feedback and requested changes on the wireframe provocations of Workshop 3 were thematically analysed by one co-lead, and then validated by the second co-lead.

4 Results

We present the results of a qualitative analysis of the co-design workshops to address 1) how end-users and decision subjects can be involved in the AI auditing process in Section 4.1, and 2) how to support auditors at each stage of a participatory audit in Section 4.2. Codebooks of the thematic analyses can be found in the corresponding sections of the Appendix. Themes are highlighted in **bold font** throughout the result section.

4.1 Involvement In Participatory Auditing

4.1.1 What Information Participatory Auditors Need. We tackled what information participatory auditors would need to support an audit, drawing on data from Workshop 1 and Workshop 3, see Appendix B Table 3 for the codebook. Most of the requested information was related to the AI application to be audited: the **Application Goals** (N=22), the overall **Application Characteristics** (N=20), who was going to be the **End-Users** (N=6), and whether the **Application Grounding** was based on domain knowledge (N=1). More specifically, for these themes, participants were interested in learning about the motivations behind the application's development, the scope of its business case, the procedures followed during development, and the criteria by which the application's success could be measured.

Participants also showed interest in the **Organisational** (N=13) aspects of the AI application development and management, such as wanting to know about the tools and processes in place to review and manage its development and deployment, handle data storage and security, and define a timeline for the different stages of the development pipeline. Furthermore, they asked for information regarding the application's **Performance** (N=8), seeking to understand when an output is correct, what errors look like and how accurate and efficient the application is. They also highlighted that there should be evidence for continuous **Performance Improvement** (N=3), related to ongoing auditing to monitor this progress.

Participants also wanted to know more about the data, specifically about the **Data Source** (N=9) used in training the AI application, highlighting concerns about possible inclusivity and diversity of the data, even though the phrasing **Data Bias** (N=1) was used only once. The group of parents in Workshop 1 Parents, in particular, wanted to know more about the backgrounds and credentials of the psychologists who provided the labels for the training data, as they play a key role in establishing the ground truth against which all children using the AI application will be compared. Participants also asked for **Data Characteristics** (N=8), emphasised the importance of **Data Consistency** (N=3) and proposed various methods to verify the reliability of the application's results, such as anonymously comparing the input of different psychologists and testing the application on multiple datasets labelled by different sources.

Participants emphasised the need for contextual information regarding the development of the AI application. They noted that auditors should be given information on the **Legal** (N=5) and **Policy** (N=1) frameworks and **Standards** (N=2) within which the AI application operates, to ensure it adheres to them. Additionally, information related to **Ethics** (N=3), **Usability** (N=2) and **Sustainability** (N=2) aspects of the AI Application were considered valuable throughout the stages of the development pipeline. In addition, participants wanted information on who the **Audit Instigator** (N=1) was. Information about the application, the data, and the audit instigator, were all details that participants requested to be shown in an *AI Application Overview* screen in an auditing tool.

4.1.2 When And Why Should An Audit Take Place. We analysed when and why an audit should be conducted through the post-its placed on the development stages, see codebook in Appendix B Table 4. Conducting an audit was most frequently suggested during the *Data Preparation* stage, followed by the *Business Case* stage. This indicates that participants across workshops agreed that AI auditing should be conducted throughout the AI development pipeline, ideally as early as possible, even when only a business case is available or when data is collected to build an AI model underlying an AI application. This confirms previous research [59], which has shown that people without AI expertise are interested in being involved in shaping an AI application in these stages, even though this is not common practice at the moment. Auditing was also often suggested during the *Testing and Evaluation* stage and in *Deployment*, which is when auditing is most commonly carried out currently.

When analysing the reasons given for auditing, participants mainly emphasised the importance of ensuring **Data Suitability** (N=10) for its intended purpose and on one occasion that it is collected through appropriate **Data Collection Processes** (N=1). Participants also expressed the need to **Evaluate the Intended Purpose** of the application (N=8), to **Maintain the Quality** (N=2) of the overall intention, and to ensure there are **No Ethical Concerns** (N=1). Furthermore, they highlighted the value of **Prevention** (N=2), starting auditing early to prevent problems later in the development pipeline.

Moreover, it was stated that auditing promotes **Accountability** (N=2) and supports the management of scope and planned data use. Participants made a case for auditing to ensure compliance with relevant **Regulations and Standards** (N=1), and to carry out any necessary **Technical Updates** (N=4) to the AI application or its environment. In addition, participants in Workshop 1 Patients mentioned that **Legal and Policy** (N=3) should be incorporated. They also were interested in monitoring the application's **Usability** (N=1).

4.1.3 Who Should Be Involved. Next, we analysed participatory auditor groups offered by participants, see codebook in Appendix B Table 5. Participants placed high importance on individuals without AI expertise being involved as auditors, such as **Domain Experts** (N=26), **Stakeholders** (N=9), **End-Users** (N=6), members of the **General Public** (N=2), or **Decision Subjects** (N=1). Especially participants in Workshop 1 Teachers and Workshop 1 Parents frequently mentioned these Domain Experts, individuals with expertise in a domain relevant to the task of the AI application, for example, child psychologists or educators in the case of *SAM*. Notably, participants in Workshop 1 Patients did not express this theme at all, possibly because this activity was done on a generic predictive AI application before we introduced *SPARRA*, which makes it harder to envision domain experts.

A further frequent response by participants across workshops was people with technical AI expertise, such as **Internal Auditors** (N=8), who in some capacity had been involved in the development of the AI application and could assist as auditors at the end of each development stage. For example, participants mentioned that development team members should be involved in the testing and evaluation stage for *SPARRA*, or that psychologists who participated in the data collection for *SAM* should be involved. Furthermore, participants stated that **AI Experts** (N=4) should be involved, however, these experts should not be affiliated with the application development team but should join audits instead as **Independent AI Experts** (N=3) to assess the integrity of the application without bias.

Other suggested auditors were **Regulatory Bodies** (N=8), such as government agencies, **Audit Experts** (N=3), **Investors** (N=3) of the AI applications, **NGOs** (N=1), or **Potential Collaborators** (N=1) who might want to work with the AI application in future. The participants of Workshop 1 Teachers highlighted the need for including a range of people to audit the AI application's **Usage Environment** (N=3) to discuss how it fits into an existing ecosystem.

4.2 Supporting Auditors During The Participatory Auditing Process

4.2.1 Impact Discovery. Overall, we have identified eight themes for Impact Discovery across the consolidated events of all three

Workshops 1. We will only present generic patterns here, but all details can be found in the codebooks in Appendix C Table 6.

The theme with the highest frequency was **Positive Impact on Care** (N=6), showing that participants were keen to discover positive outcomes. However, positive events were only discussed in workshops about *SAM*, as we only focused on negative events for *SPARRA*. While this one positive theme incorporated all positive events, there was a plethora of themes around negative events.

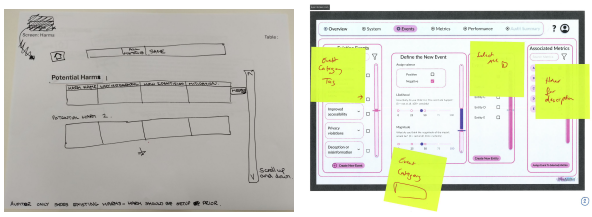
The biggest concern participants had was **Misuse** (N=4) of AI application outputs, for example, by unauthorised access. This was followed by any **Negative Impact on Care** (N=3), for example, by providing wrong support due to overreliance on AI application outputs, strain on **Resources** (N=3) due to identification of more people in need of care, and **Societal Biases** (N=3) uncovered by the use of the AI application. In addition, participants were concerned about different types of **Application Limitations** (N=3), such as the data used for development, factors that the application might not have captured, and the procedure needed to use the application. **Privacy** (N=2) concerns were also expressed, and **Representation Bias** (N=1) related to potential broader consequences of the introduction of the AI application.

There were five events that were given the highest priority in terms of likelihood and magnitude (as Workshop 2 Teachers prioritised three events with highest priority): three negative events classified in different themes (Societal Biases, Negative Impact on Care, and Resources), and two positive events.

We also analysed the events by mapping them to the taxonomy by Weidinger et al. [62]. However, classifying positive events to a harm taxonomy was not straightforward. Therefore, we decided to put a positive twist on one subcategory and created *Allocational or representational benefit* instead of harm. In addition, we noted that many events were assigned to more than one subcategory, yet two events could not be assigned to any subcategory. The events with highest priority covered all taxonomy categories except for *Human Computer Interaction Harms*. Overall, most negative events were assigned to the *Discrimination, Hate Speech and Exclusion* category (N=18), followed by *Misinformation Harms* (N=10), *Malicious Uses* (N=10), and *Information Hazards* (N=7). *Environmental and Socioeconomic harms* (N=3) and *Human-Computer Interaction Harms* (N=2) were chosen less often.

In addition, participants in Workshop 1 Parents mapped their own events to the taxonomy themselves. They created two additional events around Privacy and Misuse after the taxonomy was discussed and created a new subcategory for the *Misinformation Harms* category of the taxonomy: *potentially misleading to future research/aggregate data* to map another negative event that could not be directly assigned to the existing taxonomy. Overall, participants of this workshop assigned one event to each category.

An example of how impact discovery could be supported in an AI auditing tool with an *Event* screen can be seen in Figure 7(a), which includes information participants requested to see about events, including the valence (positive/negative), likelihood, and magnitude of the impact. Feedback by participants on the provotypes in Workshop 3 Teachers included the request for Event Categories to allow event filtering, see Figure 7(b).



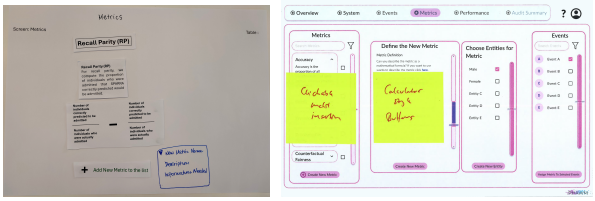
(a) Event sketch example from Workshop 3 Patients. **(b) Event Provotype from Workshop 3 Teachers.**

Figure 7: Event Screens from Workshop 3 Patients and 3 Teachers. The sketch shows a scrollable list of harms with several boxes for information, while the provotype shows an event list of existing events on the left and a method to add new events in three boxes with different event information on the rest of the screen, similar to the metric screen in Figure 6(b). Yellow post-its represent the participant feedback. See larger versions in Appendix E Figure 16.

4.2.2 Standards Identification. During Workshop 2 Patients, participants explored different strategies, see codebook in Appendix D Table 10, to formulate metrics for the events they had identified in Workshop 1 Patients. Different groups adopted different approaches to this task, while, in many cases, participants proposed applying some of the established metrics introduced earlier in the same workshop. One group began by **Defining Event Aspects** (N=4), then went on **Considering Reasons** (N=7) that were conducive to its occurrence and finally suggested a way for **Comparing the Defined Event Aspect** (N=1), followed by **Defining a Measurement Timeline** (N=2). This approach focused on discovering relevant aspects that could potentially be used to measure the event and discussing factors that can contribute to the event happening, without spending much time considering methods that provide quantifiable evidence.

By contrast, another group of participants adopted a more quantitative approach from the outset. They began by drawing a **Comparison between Prediction and Reality** (N=1), followed by a **Refinement of Groups** (N=2) until it became a mathematically formulated metric. This method focused on confirming the occurrence of the event through quantifiable evidence. However, such mathematically rigorous approaches were less common, likely reflecting participants' general discomfort with the technical aspects of metric formulation. Another challenge emerged in participants' tendency to shift focus from metric development to mitigation strategies, where discussions gravitated toward **Proposing Mitigation Methods** (N=13), rather than identifying concrete ways to measure the event itself.

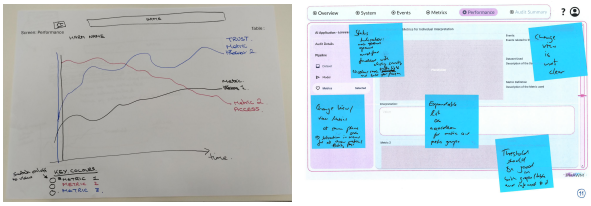
Across all Workshops 2, participants' approaches can be thematically categorised into four main types, see codebook in Appendix D Table 11: **Discovery** (N=26) of relevant aspects that could be used to measure the event, putting measures in place as **Foundation** (N=24) to be used later to measure if the event has occurred, measuring quantifiable **Evidence** (N=13) indicating the event has occurred, and a **Comparison** (N=26) between different



(a) Metrics sketch example from Workshop 3 Patients.

(b) Metrics Provotype from Workshop 3 Teachers.

Figure 8: Metrics Screens from Workshop 3 Patients and 3 Teachers. The sketch shows an example where participants chose pre-prepared information of one metric in list format, which is mirrored in the middle of the provotype (which shows the same screen which can be seen in Figure 6(b)), with green post-its showing the teacher's provided feedback. See larger versions in Appendix E Figure 17.



(a) Performance Analysis sketch example from Workshop 3 Patients.

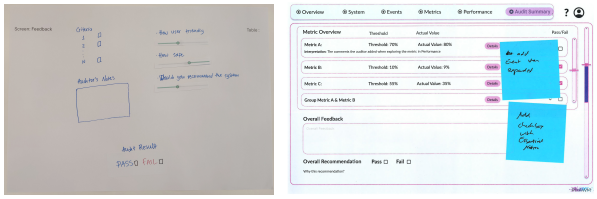
(b) Performance Analysis Provotype from Workshop 3 Parents.

Figure 9: Performance Analysis Screens from Workshop 3 Patients and 3 Parents. The sketch shows a visual presentation of the metrics performance as line graphs, while the provotype focused on the scaffolding around graphs, which were indicated with empty placeholders. Blue post-its show feedback from the parents on the screen. See larger versions in Appendix E Figure 18.

measurements to confirm the occurrence of the event. Each attempt concluded with a proposed **Threshold** (N=10) for the metric, which varied between individuals. **Contributing Factors** (N=6) and **Timeline** (N=2) were only discussed in Workshop 2 Patients.

An example *Metrics* screen to support standards identification within an auditing tool can be seen in Figure 8(a). Feedback on the provotypes, see Figure 8(b), were aimed at improving the metric creation process by presenting the options for quantitative metrics more clearly.

4.2.3 Performance Analysis. The information on Performance Analysis was mostly collected through the activities of Workshop 3. Participants agreed that the main purpose of a *Performance Analysis* screen in an auditing tool was to show the results of the metrics and wanted to see a graphical representation of this, see an example in Figure 9(a), which should include a clear presentation of the thresholds in the graphs and as text, see Figure 9(b). In addition to the performance data, information on application characteristics and goals should also be presented in some form. Metrics were



(a) Audit Outcome sketch example from Workshop 3 Patients.

(b) Audit Outcome Provotype from Workshop 3 Parents.

Figure 10: Audit Outcome Screens from Workshop 3 Patients and 3 Parents. The sketch shows a list of criteria with checkboxes, a textbox for notes and a clear pass/fail audit result option, which were all taken over into the provotype. The blue post-its show feedback from the parents. See larger versions in Appendix E Figure 19.

requested to be shown in relation to the events for which they were created.

4.2.4 Audit Communication and Advocacy. The information on Audit Communication and Advocacy was mostly collected through the activities of Workshop 3. In an auditing tool, an *Audit Outcome* screen would support participatory auditors in this step by allowing them to recommend a clear and decisive outcome, as well as feedback and comments, see Figure 10(a). Participants asked for an overview of event and metric performance to support them in making the final recommendation on the audit, see Figure 10(b). Participants highlighted that they would want a third option alongside a clear pass and fail, which they named 'Conditional', reflecting that an auditing tool might not be able to provide performance results, especially in the case of an opaque-box or translucent-box audit without access to the data needed to calculate the results of the metric. Hence, participants wanted to express that they did not have enough evidence to pass or fail the audit and wanted to express a dependency on that condition.

5 Discussion

In this section, we first address limitations of our work and then discuss the results regarding the process and value of participatory auditing. Finally we assess what this means specifically for the design of participatory auditing tools.

5.1 Limitations

Our work is not without some shortcomings that limit the generalisation of our results. Firstly, co-design workshops practically limit the number of participants, however this encourages deep and repeated engagement. While we managed to recruit 12 participants for *SPARRA*, we had difficulties recruiting teachers and parents for *SAM* due to competing time demands. Similarly, because of low participant numbers for co-design workshops, it is difficult to represent the full range of potential stakeholders adequately. We endeavoured to do this for *SPARRA*, but could not vary demographics for *SAM*. Hence, this could limit the range of perspectives we encountered in our research.

Secondly, the iterative approach of the set-up of the co-design workshops means that, unlike in controlled experimental studies, not all activities were conducted in the exact same way, so results cannot be compared directly. However, through qualitative analysis and interpretation we captured all the nuances of our data. Further evaluation studies on tools used in our research could help validate our findings.

Thirdly, we relied heavily on artefacts created by participants as the data for our analysis, only supplemented by recordings where necessary. This limited what we could analyse and how; it was much harder for participants to write down their views while also stating them. However, this also provided an 'audit trail' through our data and served to communicate individuals' perspectives to each other. Indeed, a similar approach is used in 'brainwriting', a brainstorming technique where individuals write down one idea per post-it, and then share with each other as a group.

Finally, we employed two different AI applications as grounding examples for our investigation. Thus, the specific details of what information was, or needed to be, made available diverge; however, our analysis focused on general aspects that underpin participatory auditing as a whole. AI applications and their auditing embody complex concepts and tasks, which demand time and effort to understand and carry out. Even though we carefully balanced the details of explanations and activities given the workshops' length in order not to overwhelm participants, the information provided was still dense.

5.2 Implications for Participatory Auditing Involving Stakeholders

Our findings have five main implications for establishing novel participatory auditing practices.

1) Extending auditing roles beyond the current scope. While AI experts internal to the development team might be widely seen as obvious stakeholders to assess an AI application, we found that external parties - other AI experts and regulatory bodies - were considered equally important. Clearly, our results support a shift away from *internal audits*, conducted by or on behalf of the organisation itself, to *external audits*, conducted outside of the organisation.

Furthermore, our participants clearly stated that they wanted audits to be undertaken by a broad group of participatory auditors with diverse backgrounds. Our findings strongly support the direct inclusion of decision subjects and end-users as external auditors, as suggested by Birhane et al. [6], to address power asymmetries in auditing and similar to other calls to include target/impacted communities throughout the development lifecycle of an AI application in order to achieve more equitable outcomes [20, 59].

Our results also pointed out that domain experts in the area of the AI applications, i.e., teachers in the case of *SAM*, could have an important role in participatory auditing. They provide different perspectives on potential and hidden impacts of AI applications, which may not be obvious or detectable to AI experts due to lack of deep subject expertise. An example would be the importance teachers put on the family background of the children as critical information on Data Suitability (see Appendix B Table 4), which might have been easily overlooked without their domain expertise. Involvement of domain experts can, therefore, play a critical role in auditing to uncover blindspots and areas that would be overlooked by AI experts or decision subjects on the impact of AI applications in particular contexts.

2) Refocus on auditing AI applications instead of AI models. Existing AI auditing is typically concerned with investigating AI models but neglects their impact when integrated into applications or for specific uses [6]. This restricts auditing to a purely technical evaluation of accuracy and fairness. Instead, the participants in our workshops also included *socio-technical impacts* that can only be audited in a particular application context in which the AI model is embedded. Hence, current efforts that focus on the establishment of benchmark datasets or risk taxonomies across models without consideration of the context of the AI are incomprehensive and lack value. Instead, each AI use case and application merits its own audit. This is in line with recommendations from Howell et al. [32], who make the case for reflective design in participatory auditing, as a way to elicit critical thinking that will in turn surface issues that go beyond accuracy and fairness errors, into broader ethical implications that concern the AI application as a whole.

3) Expand audits beyond current timeframe and scope. Participatory auditing should not be restricted to one particular stage of the development pipeline. Contrary to current practices, which relegate auditing to once a model has been developed or an AI application deployed, participants wanted to audit an AI application early, even when it is just an idea being conceived, and especially during data preparation that underpins the development of an AI model. This allows participatory auditors to go beyond a purely technical evaluation of the AI model and consider the socio-technical impacts [6, 23] of the AI application from the start. We argue that participatory auditing at such early stages is critical to the development of *responsible AI*, and is missing in current practices [35, 36]. What is needed are incentives for AI developers and operators, who run these AI applications, to push for greater stakeholder involvement. We believe that the results of participatory auditing can be of value to *foster trustworthiness by increasing accountability*.

4) Extend participatory auditing to cover generative AI. Our research focused on predictive AI applications in health. However, many AI applications are now built upon generative AI, such as Large Language Models (LLMs) and text-to-image AI models. Current auditing processes employed for generative AI make extensive use of *red-teaming*, where teams of AI experts "*adopt adversarial methods to identify flaws and vulnerabilities, such as harmful or discriminatory outputs from an AI system, unforeseen or undesirable system behaviours, limitations, or potential risks associated with the misuse of the system*", and the term encompasses a large variety of methods, threat modelling and follow-up activities [27]. How to extend participatory auditing to generative AI applications is still an under-studied topic, but our findings contribute to the discussion on how to support decision subjects and end-users meaningfully in this process.

5) Support the entire participatory auditing process. Birhane et al. [6] characterised an audit process which we translated into different stages of impact discovery, standards identification, performance analysis and audit communication. Currently, most tools only support specific stages, like performance analysis [38], and

often only on specific aspects such as fairness [46, 47]. Instead, we highlight the need to offer tool support for the whole process in a joined-up, end-to-end approach, from impact discovery to audit communication, which can be used by people without technical AI expertise or an auditing background.

5.3 Implications for the Design of Participatory Auditing Tools

The results of our investigations provide concrete guidelines and requirements for the design of future tools to support participatory auditing.

1) Provide information on AI applications to be audited in an accessible format. Our participants requested a substantial amount of information to support participatory audits: going beyond technical details such as AI model accuracy and fairness; they wanted to know details of the AI application's purpose, its target users, sources and diversity of data both for training as well as deployment, relevant regulation, standards and policy, and ethical considerations, usability and sustainability. Some of this information could be sourced from model cards [44], but it would need to be made usable by, and accessible to, participatory auditors without a technical background. Thus, key concepts will have to be explained in depth but in an easily understandable way, as well as supplemented by information that goes beyond models and into outlining full AI applications.

2) Support assessment of both positive and negative impacts and their effects. While auditing of predictive AI applications is often focused on potential negative outcomes or consequences, as seen by auditing stage *harms* discovery [6] and the prevalence of *harm* taxonomies for AI applications [62], our participants wanted to be able to include positive aspects in the auditing process to be able to provide a balanced view of the appropriateness of an AI application. Indeed, many of the identified events, including two of the most prioritised, were positive. Participants were also able to easily prioritise the events according to magnitude and likelihood. A tool for participatory auditing should not only allow input for positive and negative events, but should also support auditors in prioritising them in some form (which could include likelihood and magnitude) to decide if the positives of an AI application might outweigh the negatives. This would allow for a holistic audit and help participatory auditors to consider all implications of an AI application critically and recommend decisions on the appropriateness of the AI application.

3) Provide taxonomies as thinking tools, not restrictive labels. Not all created events, not even all negative ones, could easily be categorised using the taxonomy by Weidinger et al. [62], which shows that participatory auditing can broaden the typical understanding of implications of AI applications, highlighting and confirming previous findings about the usefulness of participatory auditing [23, 55]. The use of the taxonomy did encourage our participants to create additional events, confirming that the use of a taxonomy can support participatory auditors during the auditing process, as suggested by previous work [37, 60]. Participants could classify most of their own events with the help of the taxonomy, but felt the need to create additional subcategories. Therefore, if taxonomies are used in an auditing tool, the categories should function as thinking aids, allowing for expansion, and classification of positive events.

4) Design new ways to measure performance against impacts as a basis for auditing. While discussing potential impacts of the AI applications was easy for participants, they struggled with defining metrics to measure these events. Even though the adjustments made between workshops helped streamline the metric creation process, the task remained challenging for participants and highlighted the inherent difficulties of undertaking such work. This underscored the need for clear, step-by-step guidance to support participatory auditors in developing their metric ideas and help them overcome any discomfort caused by the technical nature of the task. The calculation of most metrics defined by our participants was not always immediately feasible, as they relied on variables that were either unavailable or unlikely to be attainable in a future formal audit scenario (although it should be possible for participatory auditors to require or demand certain variables to be made available during an audit). In addition to quantitative metrics, participatory auditors wanted to create qualitative metrics, which existing auditing tools relying on calculation would struggle to accommodate competently. An auditing tool will thus need to provide effective scaffolding during the metric creation process for participatory auditors. While previous work has begun to investigate this for fairness [46], further research is required to explore ways of achieving this.

5) Auditing the performance of an AI application. Visual components in the presentation of metric outcomes was universally requested by participants, with a threshold to allow for easy comparison and showing both events and the metrics they are measuring. This aligns with the recommendations by Li et al. [40], who emphasise that audit tools should facilitate the comparison of multiple pieces of evidence simultaneously, enabling users to more easily identify patterns and draw informed conclusions. These visualisations would depend on the availability of the model and data of the AI application in a transparent-box audit, without which a direct calculation of performance may not be possible. Being aware of this, participants introduced the idea of a *conditional* audit outcome in addition to clear pass and fail. This addition could also be beneficial in early auditing stages, where data might not yet be available and metrics might be of a more qualitative nature. Participants also wanted to make their final recommendation while seeing an overview of their decision for each event and metric, and wanted to be able to provide notes and feedback, so auditing tools should take this into account. Given the varying levels of model transparency and maturity between the two AI applications used in our workshops (*SPARRA* represents a long-standing model, whereas *SAM* is an early-stage research prototype), it would be interesting to compare how their respective auditing outcomes would potentially differ. However, the data from our workshops are not sufficient for such a comparison, as this would require a ready-to-use auditing tool, to enable auditors to conduct a systematic audit of each application, suitable to its opacity type. Drawing on our workshop findings to inform the design and development of such a tool can facilitate this comparison in future work.

6) Engage AI instigators with audit results. Even though participatory auditors struggled with the metric creation process,

the exercise still proved beneficial in principle, as it allowed a collection of relevant ideas to be captured and recorded, offering valuable insights that could be fed back to the audit instigators (people or organisations initiating the audit), and AI developers for accountability and to mitigate potential harm, or to collect requested data for future audits. More work will be needed to investigate in what capacity audit instigators might want, or should be required, to act on information and what form it would need to take to be useful and non-tokenistic.

7) Auditing tools for applications using generative AI. While use of predictive AI is still commonplace, it is rapidly being sidelined by generative AI, such as Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) frameworks for search. Given its popularity, much of the focus in research and practice has been placed on assessing general-purpose Generative AI, such as ChatGPT and others that are publicly available. Current approaches to assessing these generative AI systems are *internal audits*, through *red-teaming* [6], and tools that allow assessment by AI experts [2], in order to put up guardrails that prevent harmful use. A novel approach to involve end-users in assessing generative AI is through *WeAudit* [21] which offers people without an AI background, similar to our stakeholders, the ability to audit the results of prompts and report potential harms of outputs. While WeAudit introduces the concept of "collective auditing" among end-users, it does not explore intersections with legal, policy, or institutional frameworks, which we contribute in our paper. In our work, we currently focus on participatory auditing for AI applications with defined and specific tasks, not general-purpose generative AI. However, we believe that the auditing process could be extended to generative AI applications, as essentially the process we adopted, inspired by Birhane et al. [6], is technology-agnostic. Similarly, we believe that suitable tools supporting this process could be developed which measure the output produced by generative AI applications.

6 Conclusions

In this work, we addressed a crucial aspect of future responsible AI development: how to support participatory AI auditing that empowers stakeholders without auditing or AI expertise to audit an AI application. To investigate this, we conducted nine co-design workshops, involving 17 participants across two health-related AI applications. Our results show that:

- Participants clearly wanted to be involved as part of a group of external auditors. Audits must not solely rely on AI developers as is currently often the case.
- Participatory audits should be carried out throughout the AI development pipeline, instead of being relegated to predeployment or, even worse, after the AI application has been introduced.
- However, these kinds of auditors need to be properly supported. Participants wanted a wide range of information and functionalities to conduct the audit: they needed information about the AI application understandable to people without AI expertise, and the ability to include positive as well as negative impacts of the AI application to be audited. Taxonomies can be useful but should not restrict what auditors can express.
- While they were able to easily specify impacts and their effects, they struggled with how to measure them. Analysing the performance of an AI application against these impacts and metrics is challenging.

Our insights show the need to provide systematic support through tools and processes for participatory auditors. Our findings point the way for a better future in which participatory AI auditing will empower all stakeholders to shape responsible AI.

Acknowledgments

This work was supported by the Engineering and Physical Sciences Research Council [grant number EP/Y009800/1], through funding from Responsible AI UK (KP0011).

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. ACM, Glasgow Scotland Uk, 1–13. https://doi.org/10.1145/3290605.3300233

[2] Ian Arawjo, Chelse Swoopes, Priyan Vaithilingam, Martin Wattenberg, and Elena L. Glassman. 2024. ChainForge: A Visual Toolkit for Prompt Engineering and LLM Hypothesis Testing. 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 304, 18 pages. https://doi.org/10.1145/3613904.3642016

[3] Joshua Attenberg, Panos Ipeirotis, and Foster Provost. 2015. Beat the Machine: Challenging Humans to Find a Predictive Model's "Unknown Unknowns". Journal of Data and Information Quality 6, 1 (March 2015), 1–17. https://doi.org/10.1145/2700832

[4] 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 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. https://doi.org/10.48550/arXiv.1810.01943 arXiv:1810.01943 [cs].

[5] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the People? Opportunities and Challenges for Participatory AI. In Equity and Access in Algorithms, Mechanisms, and Optimization. ACM, Arlington VA USA, 1–8. https://doi.org/10.1145/3551624.3555290

[6] Abeba Birhane, Ryan Steed, Victor Ojewale, Briana Vecchione, and Inioluwa Deborah Raji. 2024. AI auditing: The Broken Bus on the Road to AI Accountability. Proceedings - IEEE Conference on Safe and Trustworthy Machine Learning, SaTML 2024 (2024), 612–643. https://doi.org/10.1109/SaTML59370.2024.00037

[7] Laurens Boer and Jared Donovan. 2012. Provotypes for Participatory Innovation. In Proceedings of the Designing Interactive Systems Conference (DIS '12). ACM, New York, NY, USA, 388–397. https://doi.org/10.1145/2317956.2318014

[8] Edyta Bogucka, Marios Constantinides, Sanja Šćepanović, and Daniele Quercia. 2024. AI Design: A Responsible Artificial Intelligence Framework for Prefilling Impact Assessment Reports. IEEE Internet Computing 28, 5 (Sept. 2024), 37–45. https://doi.org/10.1109/MIC.2024.3451351 Conference Name: IEEE Internet Computing.

[9] Aikaterini Bourazeri and Simone Stumpf. 2018. Co-designing smart home technology with people with dementia or Parkinson's disease. In Proceedings of the 10th Nordic Conference on Human-Computer Interaction (Oslo, Norway) (NordiCHI '18). Association for Computing Machinery, New York, NY, USA, 609–621. https://doi.org/10.1145/3240167.3240197

[10] John Bowlby. 1979. The making and breaking of affectional bonds. Tavistock Publication (1979).

[11] Virginia Braun, Victoria Clarke, Nikki Hayfield, and Gareth Terry. 2019. Thematic analysis. Handbook of Research Methods in Health Social Sciences (2019), 843–860. https://doi.org/10.1007/978-981-10-5251-4_103

[12] Areej Buker and Alessandro Vinciarelli. 2024. Emotion Recognition for Multimodal Recognition of Attachment in School-Age Children. In International Conference on Multimodel Interaction. ACM, San Jose Costa Rica, 312–320. https://doi.org/10.1145/3678957.3685747

[13] Christopher Burr, Sophie Arana, Cassandra Gould Van Praag, Ibrahim Habli, Marten Kaas, Michael Katell, Shakir Laher, David Leslie, Steven Niederer,

Berk Ozturk, Nuala Polo, Zoe Porter, Philippa Ryan, Malvika Sharan, Jose Solis Lemus, Marina Strocchi, and Kalle Westerling. 2024. Trustworthy and Ethical Assurance of Digital Health and Healthcare: Supporting an assurance ecosystem for data-driven technologies in health and healthcare. Technical Report. Alan Turing Institute. https://doi.org/10.5281/ZENODO.10532573

[14] Zana Buçinca, Chau Minh Pham, Maurice Jakesch, Marco Tulio Ribeiro, Alexandra Olteanu, and Saleema Amershi. 2023. AHA!: Facilitating AI Impact Assessment by Generating Examples of Harms. https://doi.org/10.48550/arXiv.2306.03280 arXiv:2306.03280 [cs].

[15] A. Cabrera, W. Epperson, F. Hohman, M. Kahng, J. Morgenstern, and D. Chau. 2019. FAIRVIS: Visual Analytics for Discovering Intersectional Bias in Machine Learning. In 2019 IEEE Conference on Visual Analytics Science and Technology (VAST). IEEE Computer Society, Los Alamitos, CA, USA, 46–56. https://doi.org/10.1109/VAST47406.2019.8986948

[16] Ángel Alexander Cabrera, Abraham J. Druck, Jason I. Hong, and Adam Perer. 2021. Discovering and Validating AI Errors With Crowdsourced Failure Reports. Proceedings of the ACM on Human-Computer Interaction 5, CSCW2 (Oct. 2021), 1–22. https://doi.org/10.1145/3479569

[17] Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jérémy Scheurer, Marius Hobbhahn, Lee Sharkey, Satyapriya Krishna, Marvin Von Hagen, Silas Alberti, Alan Chan, Qinyi Sun, Michael Gerovitch, David Bau, Max Tegmark, David Krueger, and Dylan Hadfield-Menell. 2024. Black-Box Access is Insufficient for Rigorous AI Audits. In The 2024 ACM Conference on Fairness, Accountability, and Transparency. ACM, Rio de Janeiro Brazil, 2254–2272. https://doi.org/10.1145/3630106.3659037

[18] 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. ACM, Yokohama Japan, 1–17. https://doi.org/10.1145/3411764.3445308

[19] Marios Constantinides, Edyta Bogucka, Daniele Quercia, Susanna Kallio, and Mohammad Tahaei. 2024. RAI Guidelines: Method for Generating Responsible AI Guidelines Grounded in Regulations and Usable by (Non-)Technical Roles. http://arxiv.org/abs/2307.15158 arXiv:2307.15158 [cs].

[20] Alexander d'Elia, Mark Gabbay, Sarah Rodgers, Ciara Kierans, Elisa Jones, Irum Durrani, Adele Thomas, and Lucy Frith. 2022. Artificial intelligence and health inequities in primary care: a systematic scoping review and framework. Family Medicine and Community Health 10, Suppl 1 (Nov. 2022), e001670. https://doi.org/10.1136/fmch-2022-001670

[21] Wesley Hanwen Deng, Wang Claire, Howard Ziyu Han, Jason I. Hong, Kenneth Holstein, and Motahhare Eslami. 2025. WeAudit: Scaffolding User Auditors and AI Practitioners in Auditing Generative AI. https://doi.org/10.48550/arXiv.2501.01397 arXiv:2501.01397 [cs].

[22] Wesley Hanwen Deng, Boyuan Guo, Alicia Devrio, Hong Shen, Motahhare Eslami, and Kenneth Holstein. 2023. Understanding Practices, Challenges, and Opportunities for User-Engaged Algorithm Auditing in Industry Practice. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. ACM, Hamburg Germany, 1–18. https://doi.org/10.1145/3544548.3581026

[23] Alicia DeVos, Aditi Dhabalia, Hong Shen, Kenneth Holstein, and Motahhare Eslami. 2022. Toward User-Driven Algorithm Auditing: Investigating users' strategies for uncovering harmful algorithmic behavior. In CHI Conference on Human Factors in Computing Systems. ACM, New Orleans LA USA, 1–19. https://doi.org/10.1145/3491102.3517441

[24] Patrizia Di Campli San Vito, Simone Stumpf, Cari Hyde-Vaamonde, and Gefion Thuermer. 2025. Ensuring Artificial Intelligence is Safe and Trustworthy: The Need for Participatory Auditing. CHI 2025 Workshop (2025). https://kclpure.kcl.ac.uk/ws/portalfiles/portal/329453365/CHI25_WS35_STAIG_Ensuring_Artificial_Intelligence_is_Safe_and_Trustworthy_The_Need_for_Participatory_Auditing-1.pdf

[25] Maxia Dong, Wayne H Giles, Vincent J Felitti, Shanta R Dube, Janice E Williams, Daniel P Chapman, and Robert F Anda. 2004. Insights into causal pathways for ischemic heart disease: adverse childhood experiences study. Circulation 110, 13 (2004), 1761–1766.

[26] George T. Doran. 1981. There's a S.M.A.R.T. way to write management's goals and objectives. https://community.mis.temple.edu/mis0855002fall2015/files/2015/10/S.M.A.R.T-Way-Management-Review.pdf

[27] Michael Feffer, Anusha Sinha, Wesley H. Deng, Zachary C. Lipton, and Hoda Heidari. 2024. Red-Teaming for Generative AI: Silver Bullet or Security Theater? Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society 7 (Oct. 2024), 421–437. https://doi.org/10.1609/aies.v7i1.31647

[28] National Collaborating Centre for Mental Health (UK) et al. 2015. Introduction to children's attachment. National Institute for Health and Care Excellence (NICE). https://www. ncbi. nlm. nih. gov/books/NBK356196 (2015).

[29] G7 Digital & Tech Ministers. 2023. Hiroshima AI Process International Guiding Principles for All AI Actors. https://www.soumu.go.jp/hiroshimaaiprocess/pdf/document03_en.pdf

[30] Jonathan Green, Charlie Stanley, Vicky Smith, and Ruth Goldwyn. 2000. A new method of evaluating attachment representations in young school-age children: The Manchester Child Attachment Story Task. Attachment & human development 2, 1 (2000), 48–70.

[31] Viviane Herdel, Sanja Šćepanović, Edyta Bogucka, and Daniele Quercia. 2024. ExploreGen: Large Language Models for Envisioning the Uses and Risks of AI Technologies. https://doi.org/10.48550/arXiv.2407.12454 arXiv:2407.12454 [cs].

[32] Noura Howell, Watson F Hartsoe, Jacob Amin, and Vyshnavi Namani. 2024. Reflective Design for Informal Participatory Algorithm Auditing: A Case Study with Emotion AI. In Nordic Conference on Human-Computer Interaction. ACM, Uppsala Sweden, 1–17. https://doi.org/10.1145/3679318.3685411

[33] IBM. 2019. Everyday Ethics for Artificial Intelligence. https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf

[34] ISO 19011:2018 1998. Guidelines for auditing management systems (3 ed.). Standard. International Organization for Standardization, Geneva, CH.

[35] Emma Kallina, Thomas Bohné, and Jatinder Singh. 2025. Stakeholder Participation for Responsible AI Development: Disconnects Between Guidance and Current Practice. In Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25). Association for Computing Machinery, New York, NY, USA, 1060–1079. https://doi.org/10.1145/3715275.3732069

[36] Emma Kallina and Jatinder Singh. 2024. Stakeholder Involvement for Responsible AI Development: A Process Framework. In Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization. ACM, San Luis Potosi Mexico, 1–14. https://doi.org/10.1145/3689904.3694698

[37] Sara Kingsley, Jiayin Zhi, Wesley Hanwen Deng, Jaimie Lee, Sizhe Zhang, Motahhare Eslami, Kenneth Holstein, Jason I. Hong, Tianshi Li, and Hong Shen. 2024. Investigating What Factors Influence Users' Rating of Harmful Algorithmic Bias and Discrimination. Proceedings of the AAAI Conference on Human Computation and Crowdsourcing 12 (Oct. 2024), 75–85. https://doi.org/10.1609/hcomp.v12i1.31602

[38] Michelle S. Lam, Mitchell L. Gordon, Danaë Metaxa, Jeffrey T. Hancock, James A. Landay, and Michael S. Bernstein. 2022. End-User Audits: A System Empowering Communities to Lead Large-Scale Investigations of Harmful Algorithmic Behavior. Proceedings of the ACM on Human-Computer Interaction 6, CSCW2 (Nov. 2022), 1–34. https://doi.org/10.1145/3555625

[39] Michelle S. Lam, Ayush Pandit, Colin H. Kalicki, Rachit Gupta, Poonam Sahoo, and Danaë Metaxa. 2023. Sociotechnical Audits: Broadening the Algorithm Auditing Lens to Investigate Targeted Advertising. Proceedings of the ACM on Human-Computer Interaction 7, CSCW2 (Sept. 2023), 1–37. https://doi.org/10.1145/3610209

[40] Rena Li, Sara Kingsley, Chelsea Fan, Proteeti Sinha, Nora Wai, Jaimie Lee, Hong Shen, Motahhare Eslami, and Jason Hong. 2023. Participation and Division of Labor in User-Driven Algorithm Audits: How Do Everyday Users Work together to Surface Algorithmic Harms?. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. ACM, Hamburg Germany, 1–19. https://doi.org/10.1145/3544548.3582074

[41] Peter Lovenheim. 2018. The attachment effect: Exploring the powerful ways our earliest bond shapes our relationships and lives. Penguin.

[42] Ahmed Magooda, Alec Helyar, Kyle Jackson, David Sullivan, Chad Atalla, Emily Sheng, Dan Vann, Richard Edgar, Hamid Palangi, Roman Lutz, Hongliang Kong, Vincent Yun, Eslam Kamal, Federico Zarfati, Hanna Wallach, Sarah Bird, and Mei Chen. 2023. A Framework for Automated Measurement of Responsible AI Harms in Generative AI Applications. https://doi.org/10.48550/arXiv.2310.17750 arXiv:2310.17750 [cs].

[43] Lauren Kirchner Surya Jeff Larson Mattu, Julia Angwin. [n. d.]. How We Analyzed the COMPAS Recidivism Algorithm. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[44] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency. ACM, Atlanta GA USA, 220–229. https://doi.org/10.1145/3287560.3287596

[45] Jimin Mun, Wei Bin Au Yeong, Wesley Hanwen Deng, Jana Schaich Borg, and Maarten Sap. 2025. Diverse Perspectives on AI: Examining People's Acceptability and Reasoning of Possible AI Use Cases. https://doi.org/10.48550/arXiv.2502.07287 arXiv:2502.07287 [cs].

[46] Yuri Nakao, Lorenzo Strappelli, Simone Stumpf, Aisha Naseer, Daniele Regoli, and Giulia Del Gamba. 2023. Towards Responsible AI: A Design Space Exploration of Human-Centered Artificial Intelligence User Interfaces to Investigate Fairness. International Journal of Human–Computer Interaction 39, 9 (May 2023), 1762–1788. https://doi.org/10.1080/10447318.2022.2067936

[47] Yuri Nakao, Simone Stumpf, Subeida Ahmed, Aisha Naseer, and Lorenzo Strappelli. 2022. Toward Involving End-users in Interactive Human-in-the-loop AI Fairness. ACM Transactions on Interactive Intelligent Systems 12, 3 (2022), 1–30. https://doi.org/10.1145/3514258 Publisher: Association for Computing Machinery.

[48] BBC News. 2018. Amazon scrapped 'sexist AI' tool. https://www.bbc.com/news/technology-45809919 (accessed 04/09/2025).

[49] European Parliament. 2023. EU AI Act: first regulation on artificial intelligence. https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence (accessed 09/12/2024).

[50] Erasmo Purificato, Flavio Lorenzo, Francesca Fallucchi, and Ernesto William De Luca. 2023. The Use of Responsible Artificial Intelligence Techniques in the Context of Loan Approval Processes. International Journal of Human–Computer Interaction 39, 7 (April 2023), 1543–1562. https://doi.org/10.1080/10447318.2022.2081284

[51] Marc Rettig. 1994. Prototyping for tiny fingers. Commun. ACM 37, 4 (April 1994), 21–27. https://doi.org/10.1145/175276.175288

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

[53] Jeff Sauro and Jim Lewis. 2024. 12 Things to Know About Using the TAC-10 to Measure Tech Savviness. https://measuringu.com/how-to-use-the-tac/

[54] Public Health Scotland. 2023. Scottish Individuals at Risk of Readmission and Admission (SPARRA) report on the development of SPARRA version 3. https://publichealthscotland.scot/publications/scottish-individuals-at-risk-of-readmission-and-admission-sparra-report-on-the-development-of-sparra-version-3 (accessed 09/12/2024).

[55] Hong Shen, Alicia DeVos, Motahhare Eslami, and Kenneth Holstein. 2021. Everyday Algorithm Auditing: Understanding the Power of Everyday Users in Surfacing Harmful Algorithmic Behaviors. Proceedings of the ACM on Human-Computer Interaction 5, CSCW2 (Oct. 2021), 1–29. https://doi.org/10.1145/3479577

[56] Peter Slattery, Alexander K. Saeri, Emily A. C. Grundy, Jess Graham, Michael Noetel, Risto Uuk, James Dao, Soroush Pour, Stephen Casper, and Neil Thompson. 2024. The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence. arXiv:2408.12622 [cs.AI] https://arxiv.org/abs/2408.12622

[57] Jaemarie Solyst, Cindy Peng, Wesley Hanwen Deng, Praneetha Pratapa, Jessica Hammer, Amy Ogan, Jason Hong, and Motahhare Eslami. 2025. Investigating Youth AI Auditing. https://doi.org/10.48550/arXiv.2502.18576 arXiv:2502.18576 [cs].

[58] 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 The 2024 ACM Conference on Fairness, Accountability, and Transparency. ACM, Rio de Janeiro Brazil, 713–732. https://doi.org/10.1145/3630106.3658935

[59] Beatrice Vincenzi, Simone Stumpf, Alex S. Taylor, and Yuri Nakao. 2024. Lay User Involvement in Developing Human-centric Responsible AI Systems: When and How? ACM Journal on Responsible Computing 1, 2 (June 2024), 1–25. https://doi.org/10.1145/3652592

[60] Zijie J. Wang, Chinmay Kulkarni, Lauren Wilcox, Michael Terry, and Michael Madaio. 2024. Farsight: Fostering Responsible AI Awareness During AI Application Prototyping. In Proceedings of the CHI Conference on Human Factors in Computing Systems. ACM, Honolulu HI USA, 1–40. https://doi.org/10.1145/3613904.3642335

[61] Hilde Weerts, Miroslav Dudık, Richard Edgar, Adrin Jalali, Roman Lutz, and Michael Madaio. [n. d.]. Fairlearn: Assessing and Improving Fairness of AI Systems. ([n. d.]).

[62] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of Risks posed by Language Models. In 2022 ACM Conference on Fairness Accountability and Transparency. ACM, Seoul Republic of Korea, 214–229. https://doi.org/10.1145/3531146.3533088

[63] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viegas, and Jimbo Wilson. 2019. The What-If Tool: Interactive Probing of Machine Learning Models. IEEE Transactions on Visualization and Computer Graphics (2019), 1–1. https://doi.org/10.1109/TVCG.2019.2934619

[64] Jing Nathan Yan, Ziwei Gu, Hubert Lin, and Jeffrey M. Rzeszotarski. 2020. Silva: Interactively Assessing Machine Learning Fairness Using Causality. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. ACM, Honolulu HI USA, 1–13. https://doi.org/10.1145/3313831.3376447

A Pre-prepared Screens and Content Suggestions for Workshop 3

Screen (Original: *Participants*)	Content
Login	
	User Name
	Method to prove user identity
Harms:	
Events	Harm Name
	Why is this Harmful?
Metrics	
	Metric Description
	Metric Name
	What information do you need for the metric?
Performance:	
Performance Analysis	Results of the metrics that were run to evaluate the Harm

Screen (Original: *Participants*)	Content
System to Audit:	
AI Application Overview	How are the predictions used?
	How is the prediction made?
	What informs the predictions?
	Who developed the system?
	Who monitors the system?
	What predictions are made?
Feedback:	
Audit Outcome	How to determine what's good enough to pass?
	What is the result of the Audit ?
Auditor Profile	
	Extensive demographic categories
	Name, Email, phone number

Table 2: Name and Content of pre-prepared screens for the first part of Workshop 3 Patients, discussed in Section 3.2.4. If participants requested a change in screen names in later workshops, these are shown in italics.

B Participatory Auditing

Theme	Description	Example	#
Application Characteristics	Overall information about the AI Application	WS1 Patients: Privacy of data Input	20
Application Goals	What the AI Application is aiming to achieve	WS1 Parents: Review of business case purpose - diagnosis or candidates for diagnosis?	22
Application Grounding	AI Application is grounded in domain knowledge	WS1 Teachers: Knowledge of and understanding of attachment theory - professional	1
Audit Instigator	Information on who instigated the audit	WS3 Patients: Who has initiated the audit	1
Auditor with AI Experts	Information on who has expertise in auditing and AI	WS1 Patients: An knowledgeable Auditor having expertise on AI	1
Data Bias	Information about data biases	WS1 Patients: Biases of the Data (sample size, diversity)	1
Data Characteristics	Overall information of data	WS1 Patients: Data parameters	8
Data Consistency	Information about data consistency	WS1 Parents: blind test of results from other psychologists, see how they compare vs training psychologists	3
Data Source	Information about data source	WS1 Parents: Psychologist's backgrounds & credentials as they relate to the child demographics	9
End-Users	People who use the AI Application	WS1 Teachers: Target Audience: Social Work/Education/CALMs	6
Ethics	Ethics are being considered	WS1 Teachers: Safeguarding	3
Legal	Information about legal regulations	WS1 Patients: Legal framework	5
Organisational	Information about organisational processes	WS1 Patients: Who verifies / authorises /confirms / access / security.	13
Performance	Information about the performance of the AI Application	WS1 Patients: Model accuracy	8
Performance Improvement	Information about the performance improvement of the AI Application	WS1 Patients: AI system model Efficiency whether it is constantly improving.	3
Policy	Information about policy regulations	WS1 Patients: Existing policy in deployment.	1
Stakeholders	Information about who might be interested in the AI Application	WS1 Patients: Stakeholders	1
Standards	Information about standards regulations	WS1 Patients: Any standards and recommended practices that are in place.	2
Sustainability	Information about the use of resources	WS1 Patients: Evaluating the environmental impact, such as the amount of cooling water required to cool the processes and hardware.	2
Usability	Information about usability	WS1 Patients: Usability and accessibility of user interface. Interface.	2

Table 3: Codebook of the thematic analysis on what information the workshop participants considered necessary to carry out an AI audit, as discussed in Section 4.1. For each theme we show the name, description, an example citation from the workshops and how often this theme occurred (#).

Theme	Description	Example	#
Accountability	Ensure accountability	WS1 Patients: Accountability principle employed at every step of the process for the benefit of the developers knowledge.	2
Bias	AI Application or underlying data show bias	WS1 Patients: Bias issue	1
Data Collection Processes	Ensure that appropriate procedure were followed	WS1a: Check if the right processes have been adhered to/ for completeness of the processes	1
Data Security	Ensure data security	WS1.s: Security of the data	1
Data Suitability	Ensure data suitability	WS1 Teachers: Family background	10
Evaluate the Intended Purpose	Ensure that the outcomes fulfil the goals	WS1 Parents: Does step 5 match step 1? Have we achieved objectives?	8
Legal and Policy	Information about legal regulations	WS1 Patients: Violation of law/rules	3
Maintain the Quality	Ensure that quality of AI Application is maintained	WS1 Patients: For quality and consistency	2
No Ethical Concerns	Ensure ethics are being considered	WS1 Patients: Evaluate the ethics & morality	1
Organisational	Information about organisational processes	WS1 Patients: Pre-development for management of scope & planned data use. Planned	1
Prevention	To prevent adverse outcomes	WS1 Patients: Prevention is better than cure	2
Regulations and Standards	Information about standards regulations	WS1 Patients: Check if the model is compliant to any set standards of operations.	1
Technical Updates	Making technical or overall updates to the AI Application or its environment	WS1 Patients: Post-deployment. To check; human, computer interaction/change	4
Usability	Ensuring the application's usability	WS1 Patients: Is it user friendly ?	1

Table 4: Codebook of the thematic analysis on when (and why then) the workshop participants considered it necessary to carry out an AI audit, as discussed in Section 4.1. For each theme we show the name, description, an example citation from the workshops and how often this theme occurred (#).

Theme	Description	Example	#
AI Experts	People with prior AI knowledge	WS1 Patients: Programmers	4
Audit Experts	People with prior knowledge in auditing	WS1 Patients: System Auditors	3
Decision Subjects	People on whom the AI Application makes predictions	WS1 Patients: Focus group Eg for medical AI. patients with the target condition	1
Domain Experts	People with knowledge in the domain in which the AI Application operates	WS1 Teachers: Psychologists/ school nurse/ Educators/ Health	26
End-Users	People who use the AI Application	WS1 Parents: Wider professional user base	6
General Public	Anyone	WS1 Patients: A variety of people such as ethical moral people.	2
Independent AI Experts	People outside the development organisation with prior AI knowledge	WS1 Parents: Wider AI developers	3
Internal Auditors	People within the development organisation who audit the AI Application	WS1 Parents: Developers of this system continuing to validity check.	8
Investors	Entities providing finances	WS1 Patients: Shareholders those who input the Capital.	3
NGOs	Non-government organisations	WS1 Patients: NGOs	1
Potential Collaborators	Entities who might want to collaborate with the development organisation	WS1 Patients: Potential Collaborators	1
Regulatory Bodies	Entities in charge of regulation	WS1 Patients: Government Agencies	8
Stakeholders	Entities with a stake in the AI Application	WS1 Patients: Impact on people to whom it may have a positive or negative effect.	9
Usage Environment	Relating to the environment where the AI Application is going to be used	WS1 Teachers: Working with range of focus groups to plan adoptions	3

Table 5: Codebook of the thematic analysis on who the workshop participants considered should carry out an AI audit, as discussed in Section 4.1. For each theme we show the name, description, an example citation from the workshops and how often this theme occurred (#).

C Impact Discovery

Theme	Description	Example	#
Application Limitations	Data used to develop the AI Application might lead to incorrect predictions	WS1 Parents: Errors or lack of diversity in training data could lead to misidentification	3
Misuse	AI Application output being used for purposes other the intended	WS1 Patients: Exploitation of data by unauthorised third parties.	4
Negative Impact on Care	AI Application might cause healthcare support to decline	WS1 Teachers: Overreliance on results without context or safeguarding can lead to wrong support being put in place.	3
Positive Impact of Care	AI Application might cause healthcare support to improve	WS1 Teachers: Universal assessment could identify children with insecure attachment that would previously have been missed.	6
Privacy	The privacy of decision subject might be violated	WS1 Patients: Privacy and patient anonymity might be at risk.	2
Representation Bias	AI Application might introduce biases in the way the decision subjects are represented	WS1 Parents: The introduction of *SAM* could lead to misrepresentation of children with insecure attachment in research due to self-selection.	1
Resources	AI Application will lead to increase in classifications and strain available resources	WS1 Teachers: False positive classification of insecure attachment can lead to waste of human and financial resources.	3
Societal Biases	AI Application uncovers pre-existing biases in society and/or exposes decision subjects to those	WS1 Patients: Inaccurate scores for patients with low resource access (due to insufficient data) can lead to untimely diagnoses.	3

Table 6: Codebook of the thematic analysis of Events, as discussed in Section 4.2.1. For each theme we show the name, description, an example citation from the workshops and how often this theme occurred (#).

WS	Event Name	Prio	V	Theme	Taxonomy Subcategory
WS1 Patients	Inaccurate scores for patients with low resource access (due to insufficient data) can lead to untimely diagnoses.	1	N	Societal Biases	•Allocational or representational harm •Uneven performance for different social groups •Social exclusion •Increasing social inequalities from uneven distribution of risk and benefits
WS1 Patients	Other factors not captured in the model might influence the predictions.	4	N	Application Limitations	•Deceiving or misinforming a user
WS1 Patients	Reduced Quality of healthcare due to overreliance on predictions.	3	N	Negative Impact on Care	
WS1 Patients	Privacy and patient anonymity might be at risk.	5	N	Privacy	•Privacy violations •Facilitating fraud, scam, impersonation crimes •Creating avenues to exploit or violate privacy of the user
WS1 Patients	Exploitation of data by unauthorised third parties.	2	N	Misuse	•Privacy violations •Safety risks •Deceiving or misinforming a user •Facilitating fraud, scam, impersonation crimes •Personalised disinformation campaigns •Weaponisation or production of malicious code
WS1 Patients	Lack of trust by patients caused by system misuse.	6	N	Misuse	•Growing societal distrust in shared information •Undermining public discourse
WS1 Patients	Lack of trust by GPs caused by system misuse.	7	N	Misuse	•Growing societal distrust in shared information •Undermining public discourse

Table 7: Event Themes and Taxonomy Subcategory Classification for all negative WS1 Patients events, as discussed in Section 4.2.1. For each event, we show the workshop, the event name, the prioritisation ranking of the participants (Prio), the valence (V) of the event, which could only be negative (N) in this workshops as we focused on harms only, as well as the theme the event was coded into and the taxonomy subcategory the theme was assigned to (based on Weidinger et al. [62]).

WS	Event Name	Prio	V	Theme	Taxonomy Subcategory
WS1 Teachers	Children identified with insecure attachment by the system could miss out on support for other issues.	8	N	Negative Impact on Care	•Allocational or representational harm •Increasing social inequalities from uneven distribution of risk and benefits
WS1 Teachers	Universal assessment could identify children with insecure attachment that would previously have been missed.	4	P	Positive Impact of Care	•B: Allocational or representational harm
WS1 Teachers	Overreliance on results without context or safeguarding can lead to wrong support being put in place.	1	N	Negative Impact on Care	•Allocational or representational harm •Deceiving or misinforming a user
WS1 Teachers	Children identified to have insecure attachment, as well as their families, can be provided with support through educators and psychologists.	10	P	Positive Impact of Care	•B: Allocational or representational harm
WS1 Teachers	Universal assessment could identify children with insecure attachment earlier and provide feedback to parents	2	P	Positive Impact of Care	•B: Allocational or representational harm
WS1 Teachers	Parents or whole families could be stigmatised when children get labelled as having insecure attachment.	9	N	Societal Biases	•Allocational or representational harm •Profound offence or psychological harm •Social exclusion •Weaponisation or production of malicious code
WS1 Teachers	Universal use of *SAM* could lead to higher numbers of children labelled with insecure attachment, overwhelming human and financial resources.	3	N	Resources	•Allocational or representational harm •Material harm
WS1 Teachers	The system could be integrated into existing profiles for a more holistic presentation of children's wellbeing.	6	P	Positive Impact of Care	•B: Allocational or representational harm
WS1 Teachers	False positive classification of insecure attachment can lead to waste of human and financial resources.	5	N	Resources	•Allocational or representational harm •Material harm
WS1 Teachers	The labelling of children could lead to a break in relationship between parents and educators.	7	N	Societal Biases	•Profound offence or psychological harm •Social exclusion

Table 8: Event Themes and Taxonomy Subcategory Classification for all WS1 Teachers events, as discussed in Section 4.2.1. For each event, we show the workshop, the event name, the prioritisation ranking of the participants (Prio), the valence (V) of the event, which could be positive (P) or negative (N), as well as the theme the event was coded into and the taxonomy subcategory the theme was assigned to (based on Weidinger et al. [62]). *B* in front of the taxonomy subcategory indicates the group was used as benefit, i.e., instead of harm it was a benefit.

WS	Event Name	Prio	V	Theme	Taxonomy Subcategory	Taxonomy Subcategory Participants
WS1 Parents	Children identified to have insecure attachment can receive evidence for referral for further assessment and parents can improve parenting strategies and arrangements.	1	P	Positive Impact of Care	•B: Allocational or representational harm	
WS1 Parents	Use of *SAM* could lead to improved health provision efficiency and reduced waiting times for psychologists.	5	P	Positive Impact of Care	•B: Allocational or representational harm	
WS1 Parents	Children could feel disrupted by and react differently to the testing procedure.	3	N	Application Limitations		
WS1 Parents	Errors or lack of diversity in training data could lead to misidentification.	2	N	Application Limitations	•Allocational or representational harm •Uneven performance for different social groups •Social exclusion •Increasing social inequalities from uneven distribution of risk and benefits	•Uneven performance for different social groups •Perpetuating discriminatory stereotypes via product design •Increasing social inequalities from uneven distribution of risk and benefits
WS1 Parents	Labelling information could be used against individuals in later life.	4	N	Misuse	•Allocational or representational harm •Social exclusion •Privacy violations •Safety risks •Growing societal distrust in shared information •Unethical actions by users •Augment illegitimate mass surveillance •Personalised disinformation campaigns •Weaponisation or production of malicious code •Creating avenues to exploit or violate privacy of the user	•Augment illegitimate mass surveillance •Creating avenues to exploit or violate privacy of the user
WS1 Parents	Misidentification of insecure attachment or overreliance on results can lead to waste of resources.	7	N	Resources	•Allocational or representational harm •Material harm	•Deceiving or misinforming a user
WS1 Parents	The introduction of *SAM* could lead to misrepresentation of children with insecure attachment in research due to self-selection.	6	N	Representation Bias		**New Misinformation Harms subcategory:** potentially misleading to future research/aggregate data
WS1 Parents	Insufficient privacy safeguards could lead to privacy violations.	8	N	Privacy	•Privacy violations •Safety risks	•Privacy violations

Table 9: Event Themes and Taxonomy Subcategory Classification for all WS1 Parents events, as discussed in Section 4.2.1. For each event, we show the workshop, the event name, the prioritisation ranking of the participants (Prio), the valence (V) of the event, which could be positive (P) or negative (N), as well as the theme the event was coded into and the taxonomy subcategory the theme was assigned to (based on Weidinger et al. [62]). *B* in front of the taxonomy subcategory indicates the group was used as benefit, i.e., instead of harm it was a benefit. In addition, we show the taxonomy subcategories the participants themselves chose for their events, which includes a new subcategory they created in the workshop.

D Standards Identification

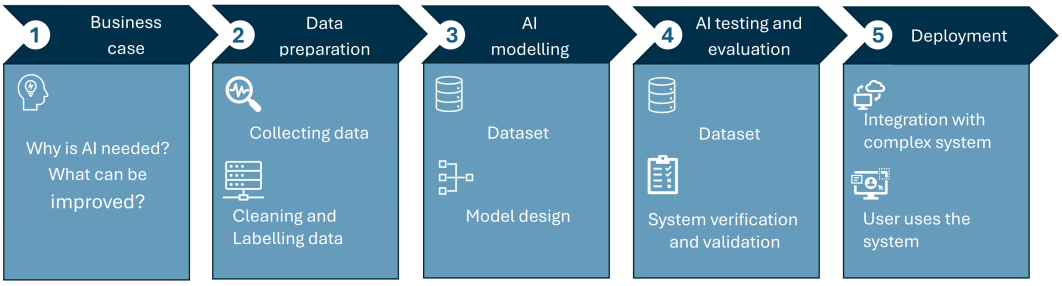
Theme	Description	Example	#
Applying Established Metric	Using an established metric that was previously discussed	The recall metric is quite useful to predict the accuracy to some extent	3
Comparing Event Related Aspects between Groups	Aspects related to the AI application but not including the AI application output are compared between groups	Disparity between Death rate at deprived areas & Death rate at high resource areas in younger demographics and patients with chronic diseases such as diabetes	1
Comparing the Defined Event Aspect	Pre-defined event aspects are compared	Metric = people with certain lower income/total number of people in that postcode x 100%	1
Comparison between Prediction and Reality	Comparing the prediction with actual numbers	Ratio of patients in known low resource area in patient admission predictions over actual inpatient admissions in the identified area	1
Considering Outcomes	Discussing potential outcomes of the event	unwanted advertising of specific goods/services	2
Considering Reasons	Discussing potential reasons why the event may occur	oversubscription to a particular GP which causes overwhelmness	7
Defining Event Aspect	Discussing an aspect of the event	Delay for access to a GP	4
Defining a Measurement Timeline	Discussing the timeline for continuous evaluation of the event	There should be a survey half yearly	2
Defining Quantifiable Effect of the Event	Discussing an aspect of the event which can be quantified	Measure by looking at the rate of death in a derived area	3
Monitoring Event Aspects	Discussing how occurrence of an aspect of the event could be monitored	Self report by respondents	4
Monitoring Quantifiable Event Aspects	Discussing how occurrence of a quantifiable aspect of the event could be monitored	Self report by respondents & Number of members within a GP with access to data to input into *SPARRA* , ie patient records	1
Proposing Mitigation Methods	Discussing how the event could be mitigated	Specific app in the mobile installed with password & IDs	13
Refinement of Groups	Pre-defined groups are refined or specified	Death rate at deprived area - Death rate at high resource area	2
Reiteration	Pre-discussed metric is reiterated without change	admission predictions in deprived/actual admissions in deprived - admission predictions in non-deprived/actual admissions in non-deprived	6

Table 10: Codebook of the thematic analysis of Metric Creation Strategies in Workshop 1 Patients, as discussed in Section 4.2.2. For each theme we show the name, description, an example citation from the workshops and how often this theme occurred (#).

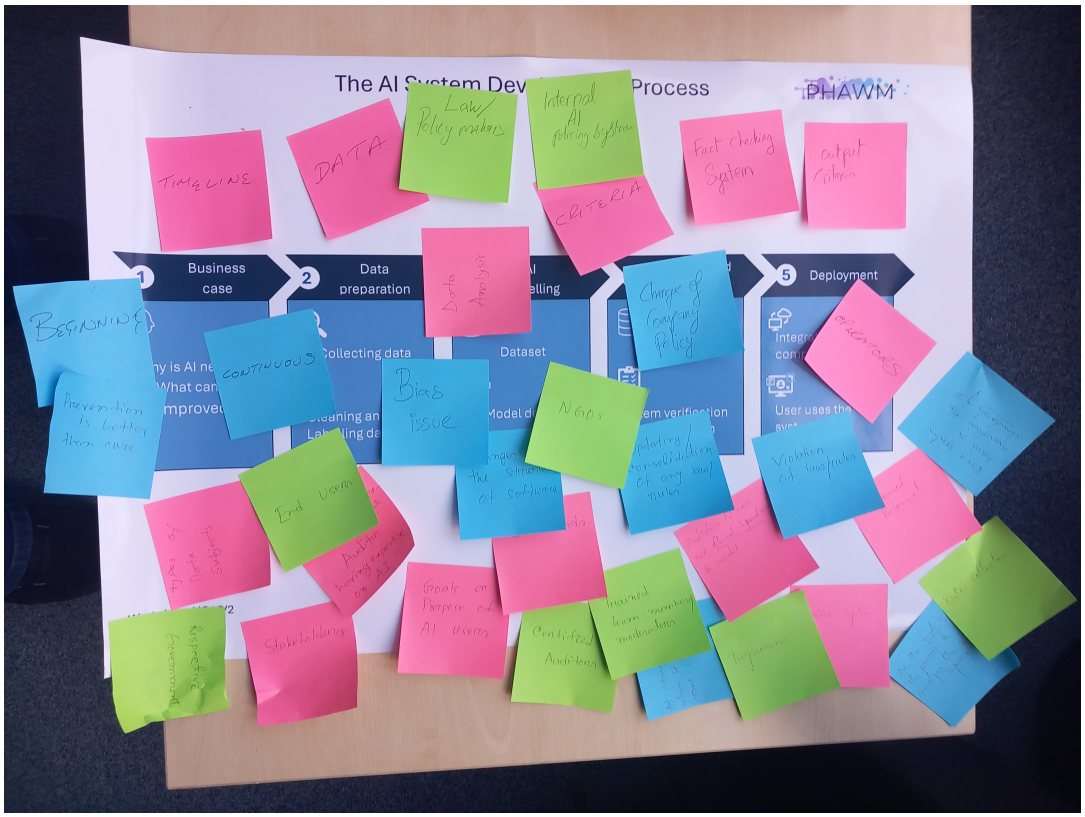
Theme	Description	Example	#
Comparison	Comparison between different measurements to confirm the event happening	WS2 Patients: Disparity between Death rate at deprived areas & Death rate at high resource areas in younger demographics and patients with chronic diseases such as diabetes	26
Contributing Factors	Factors that can contribute to the event happening	WS2 Patients: Lack of facility/inefficient facility	6
Discovery	Discovering relevant aspects that might be used to measure the event	WS2 Parents: individual stress reactions of individual children measured by body language and verbal cues	26
Evidence	Measure quantifiable evidence for an event happening	WS2 Parents: (all uses of resources manual process x,y,z) - (all uses of resources AI process a,b,c)	13
Foundation	Putting measures in place as a foundation to be used later to measure if an event happened	WS2 Teachers: cost of after care usual X number of people *SAM* classified	24
Threshold	Defining the acceptable extent of an event	WS2 Teachers: 80% correct	10
Timeline	Defining a timeline for the frequency of the measurement	WS2 Patients: Monthly review	2

Table 11: Codebook of the thematic analysis of Metric Creation Themes, as discussed in Section 4.2.2. For each theme we show the name, description, an example citation from the workshops and how often this theme occurred (#).

E Artefacts

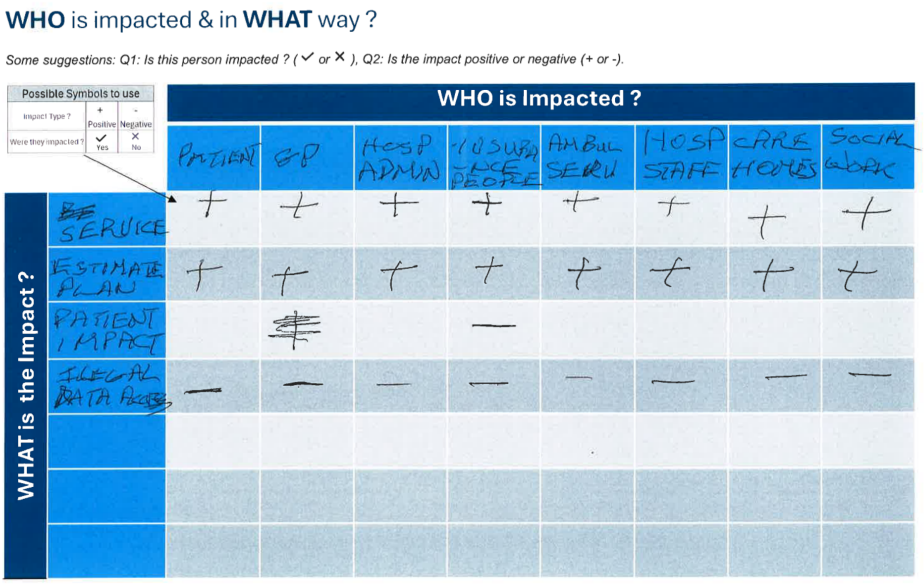


(a) AI Development Pipeline used in Workshop 1, adapted from Vincenzi et al. [59] with permission.

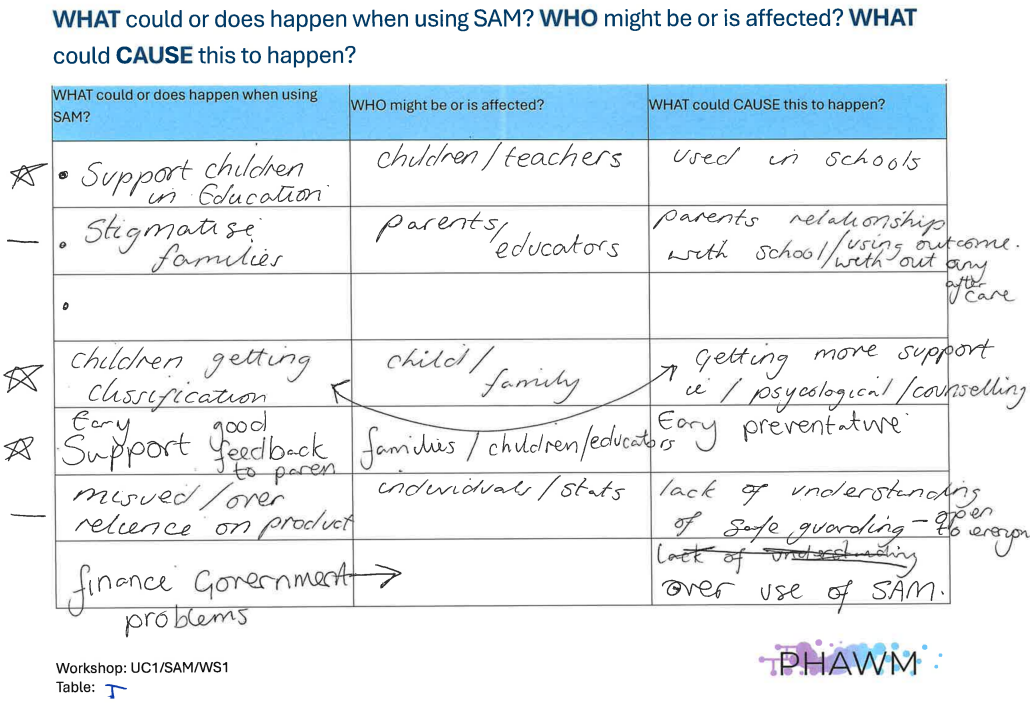


(b) AI Development Pipeline with post-it of what information was needed (pink), when the application should be audited (blue) and who should audit it (green)

Figure 11: Larger version of AI Development Pipeline with post-its and Impact Assessment Matrix from activities in Workshop 1 Patients. Relating to Section 3.2.2 Figure 3.



(a) Impact Assessment Matrix



(b) Event Table

Figure 12: Larger version of Impact Assessment Matrix from Workshop 1 Patients and Event Table from Workshop 1 Teachers. Relating to Section 3.2.2 Figure 4.

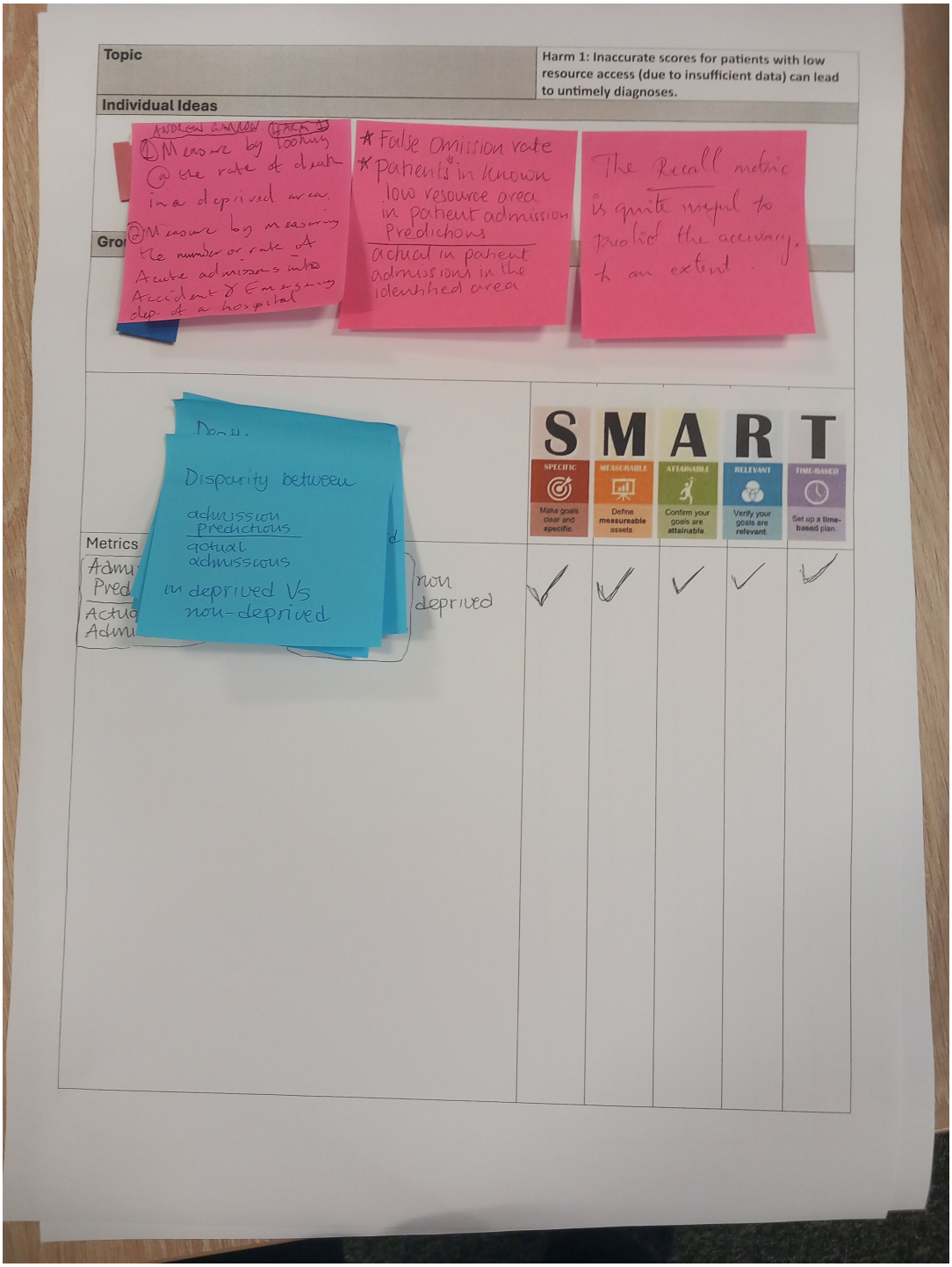
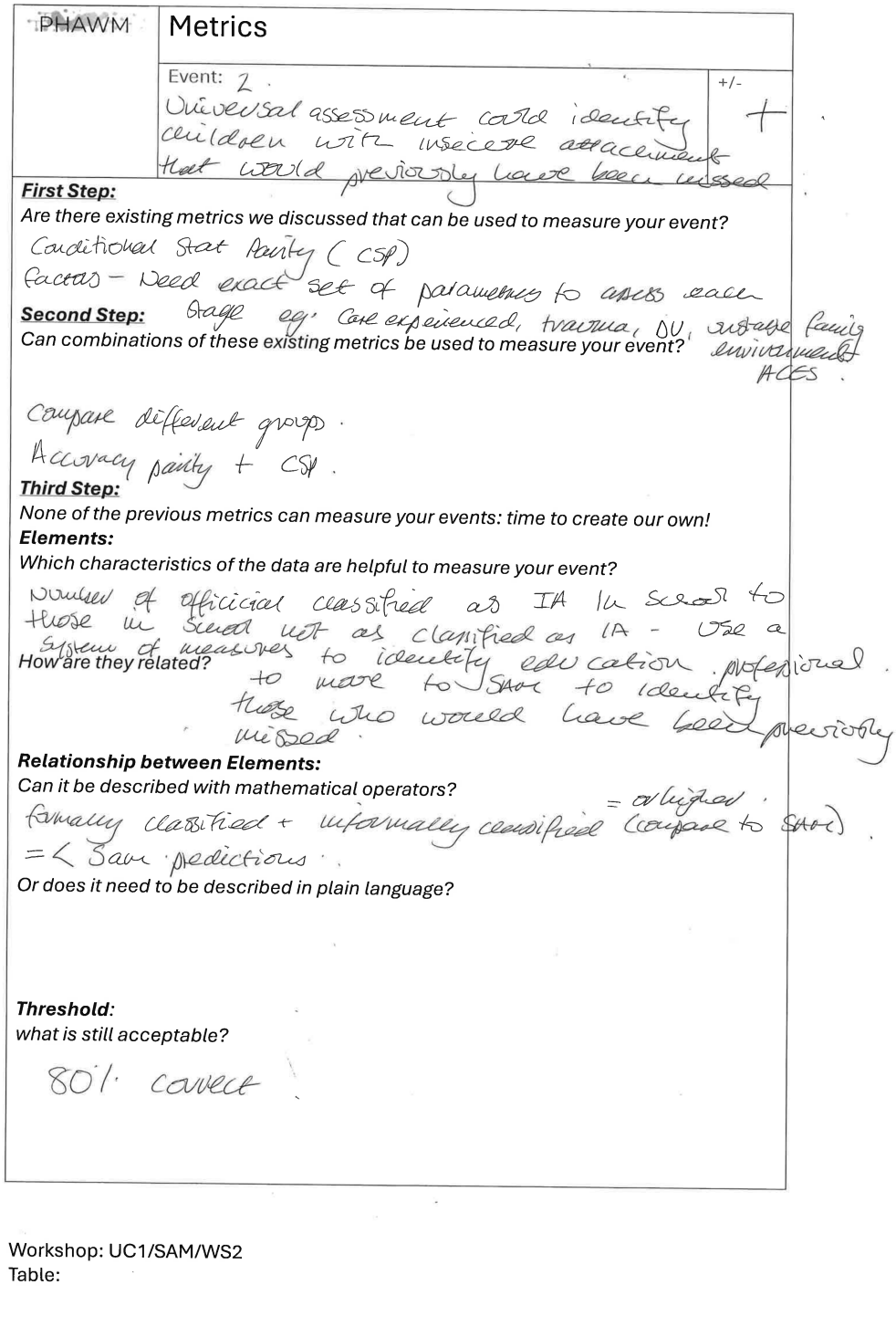
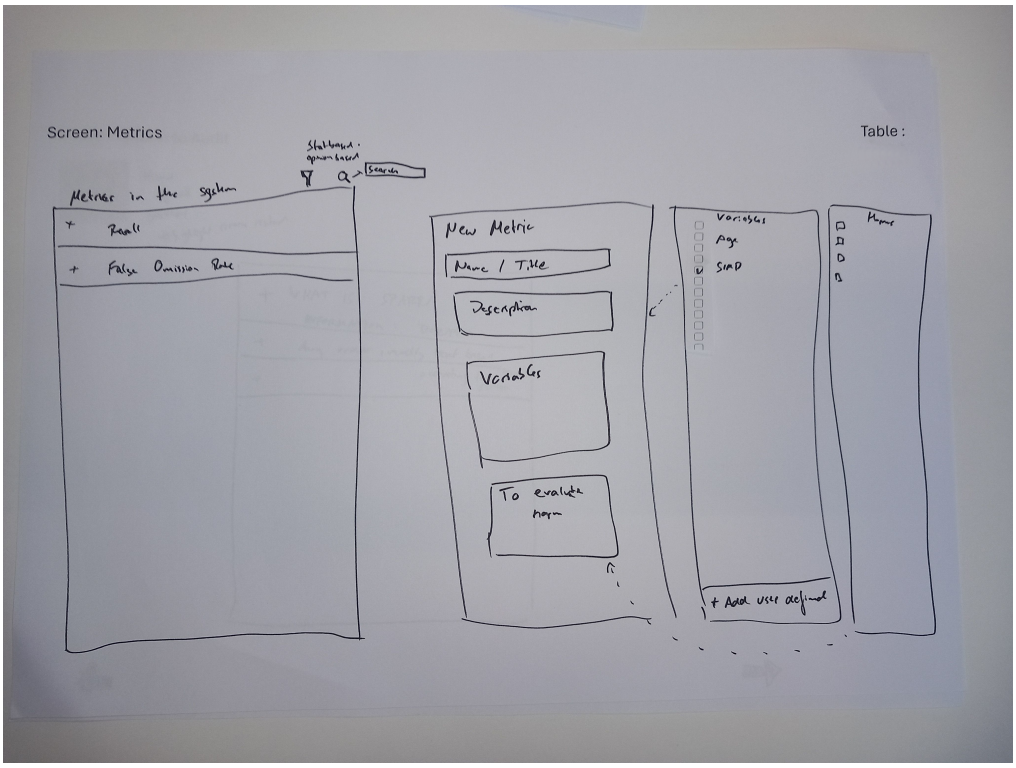


Figure 13: Larger version of Metric Creation with SMART from Workshop 2 Patients. Relating to Section 3.2.3 Figure 5(a).

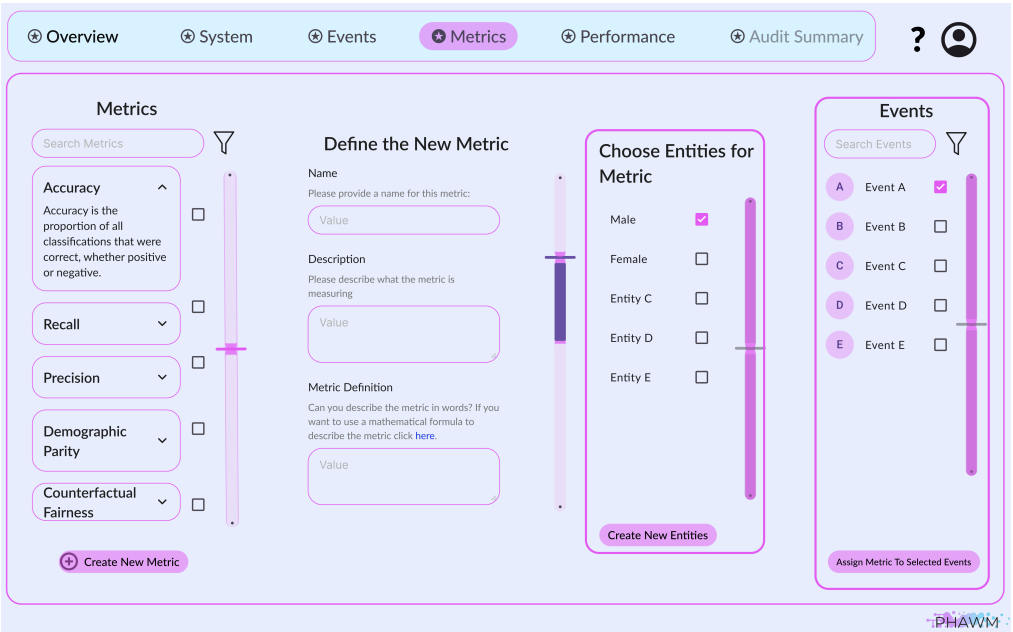


Workshop: UC1/SAM/WS2
Table:

Figure 14: Larger version of Metric Creation with three cases as steps from Workshop 2 Teachers. Relating to Section 3.2.3 Figure 5(b).

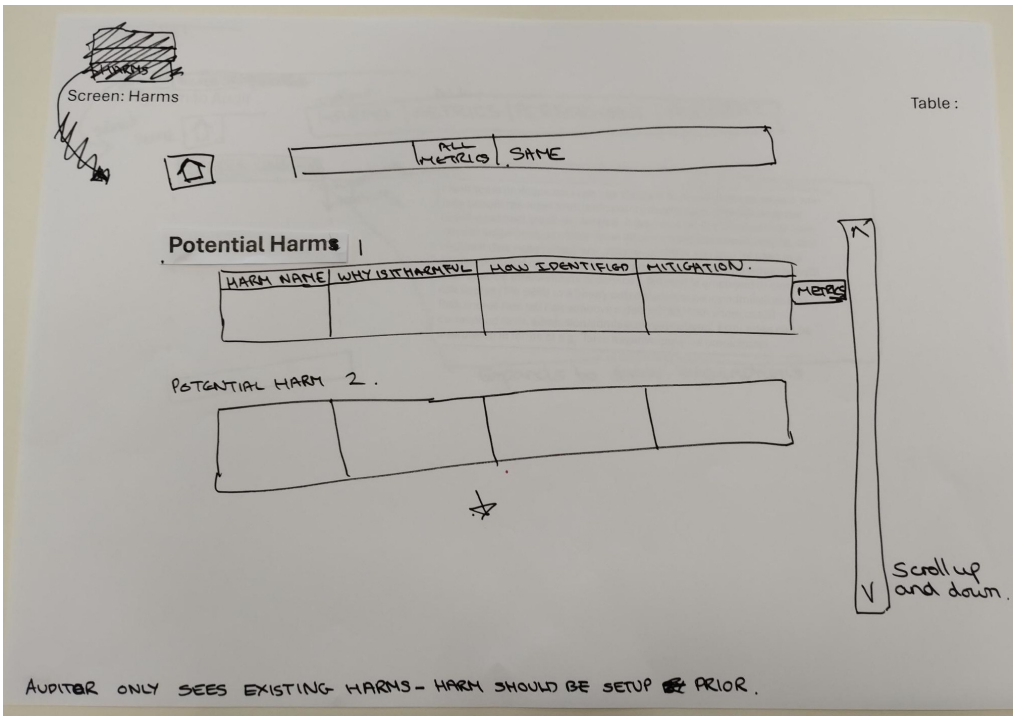


(a) Wireframe sketch from Workshop 3 Patients.



(b) Wireframe provotype from Workshop 3 Teachers/Parents.

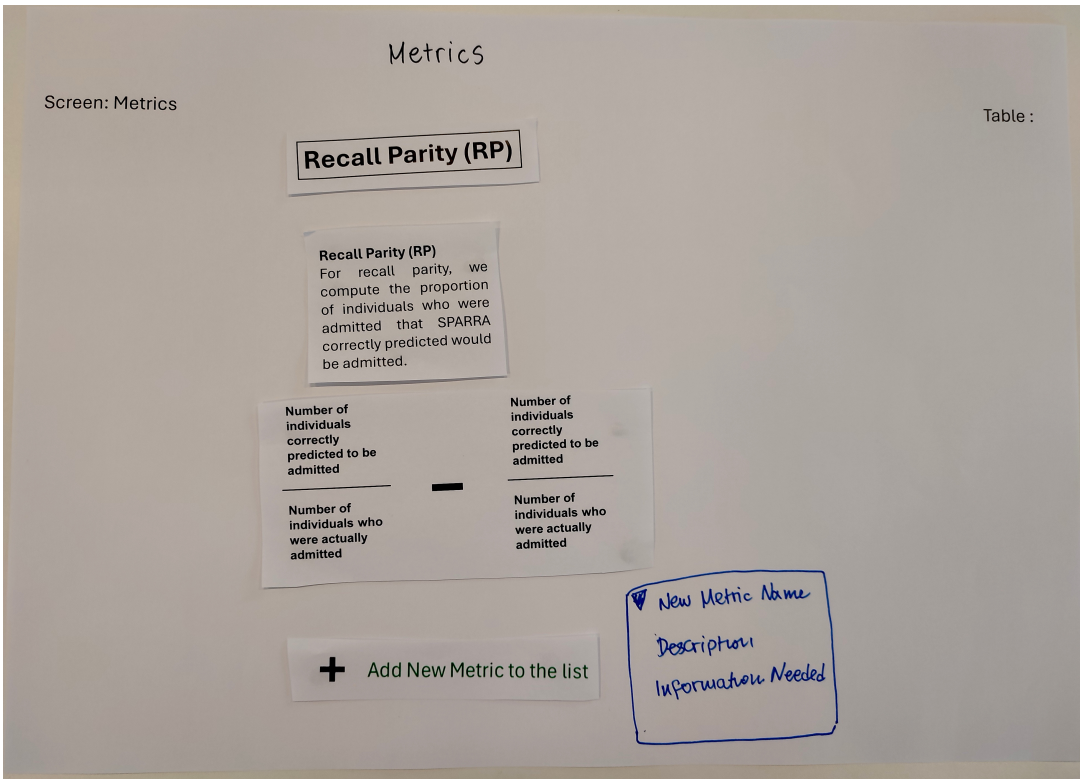
Figure 15: Larger version of Wireframe sketch from Workshop 3 Patients and wireframe provotypes used in Workshop 3 Teachers/Parents, showing the Metrics screen with a list of metrics already in the system and a mechanism to create new metrics. The main elements of the screens are the same, with only small changes between the sketch and the wireframe, such as added checkboxes for the already existing metrics. Relating to Section 3.2.4 Figure 6.



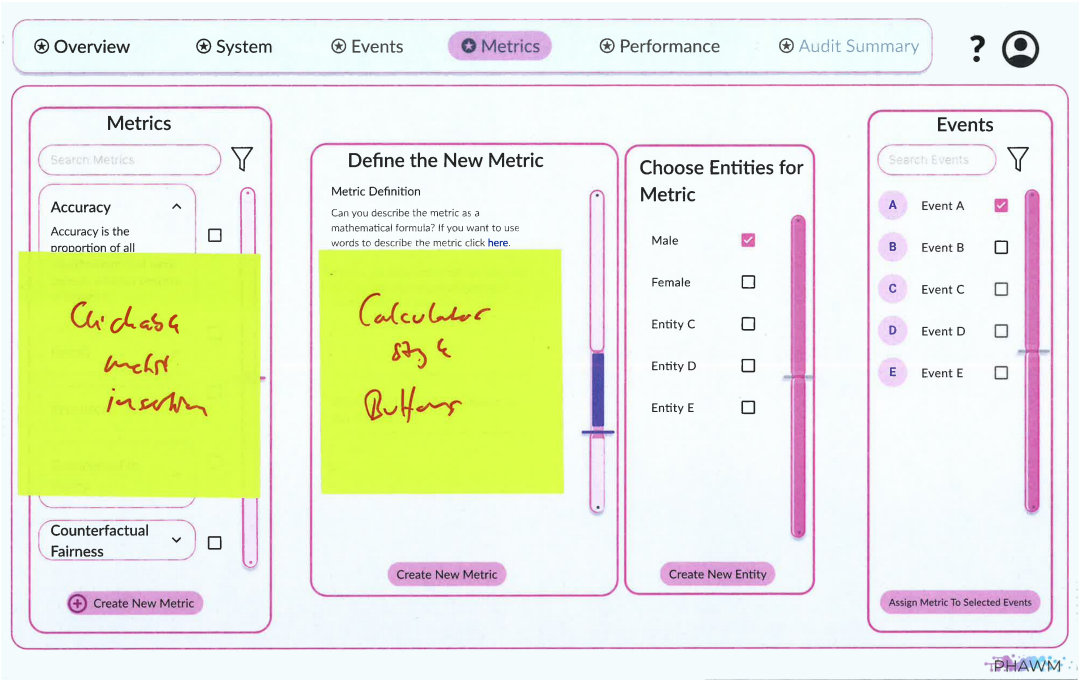
(a) Event sketch example from Workshop 3 Patients.

(b) Event Provotype from Workshop 3 Teachers.

Figure 16: Larger version of Event Screens from Workshop 3 Patients and 3 Teachers. The sketch shows a scrollable list of harms with several boxes for information, while the provotype shows an event list of existing events on the left and a method to add new events in three boxes with different event information on the rest of the screen, similar to the metric screen in Figure 6(b). Yellow post-its represent the participant feedback. Relating to Section 4.2 Figure 7.

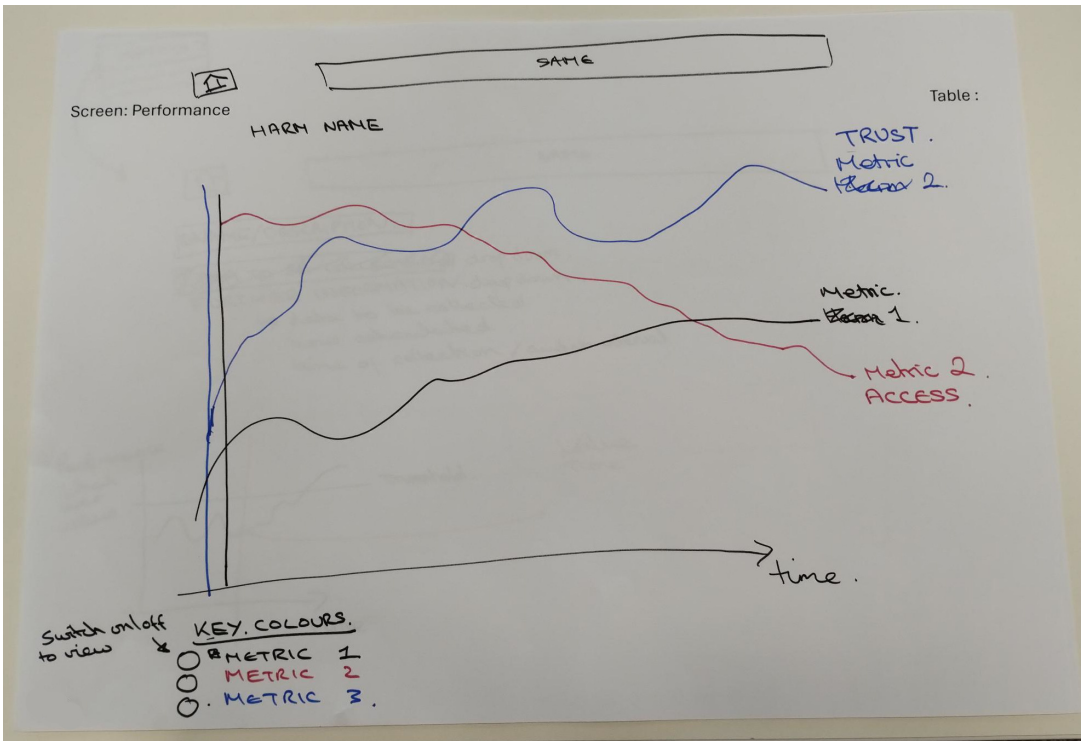


(a) Metrics sketch example from Workshop 3 Patients.

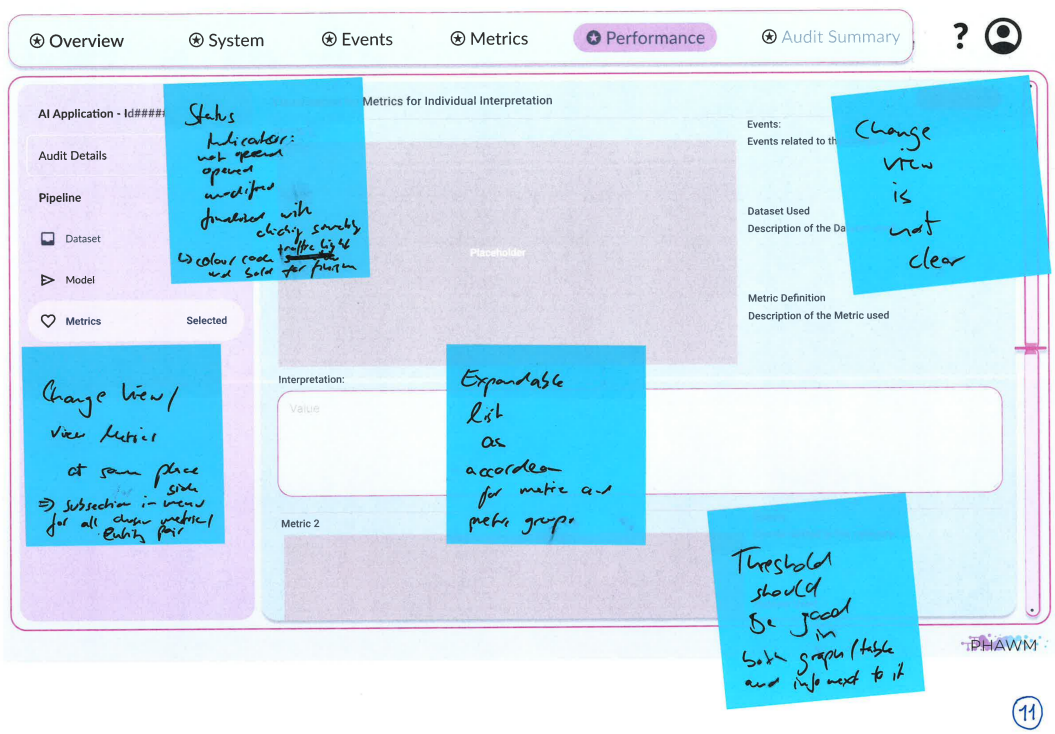


(b) Metrics Provotype from Workshop 3 Teachers.

Figure 17: Larger version of Metrics Screens from Workshop 3 Patients and 3 Teachers. The first screen shows an example where participants chose pre-prepared information of one metric in list format, which is mirrored in the second picture (which shows the same screen which can be seen in Figure 6(b)) in the middle, hidden under green post-its with which the teachers provided feedback. Relating to Section 4.2.2 Figure 8.

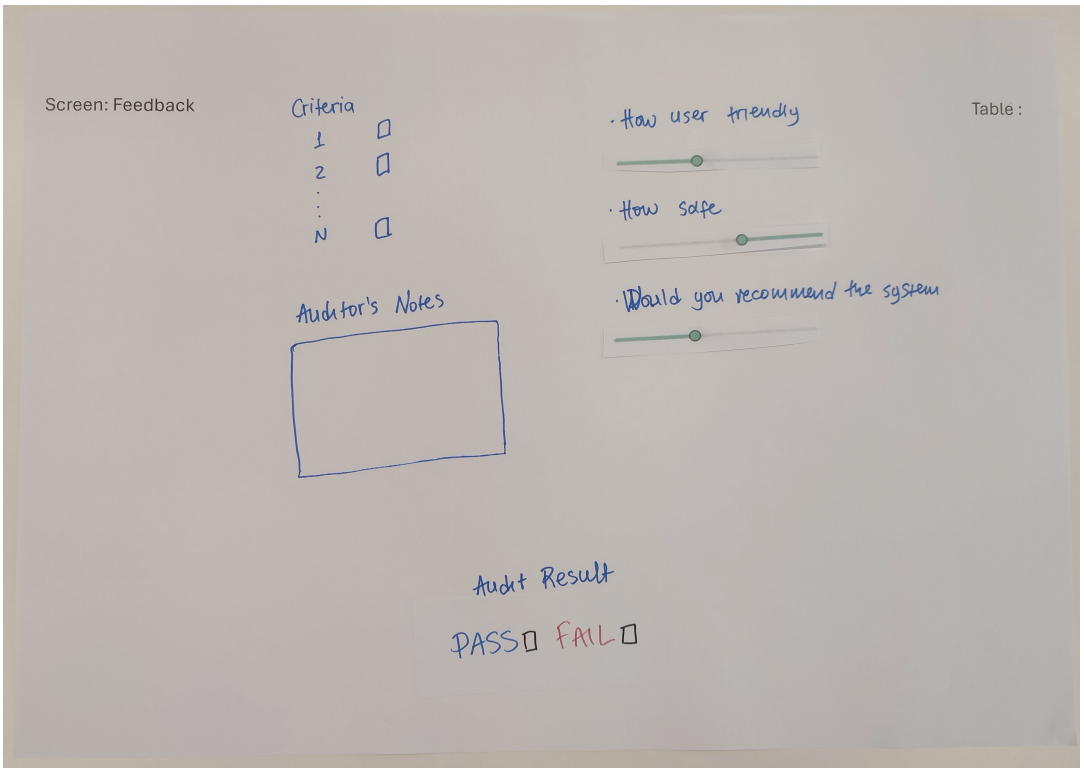


(a) Performance Analysis sketch example from Workshop 3 Patients.

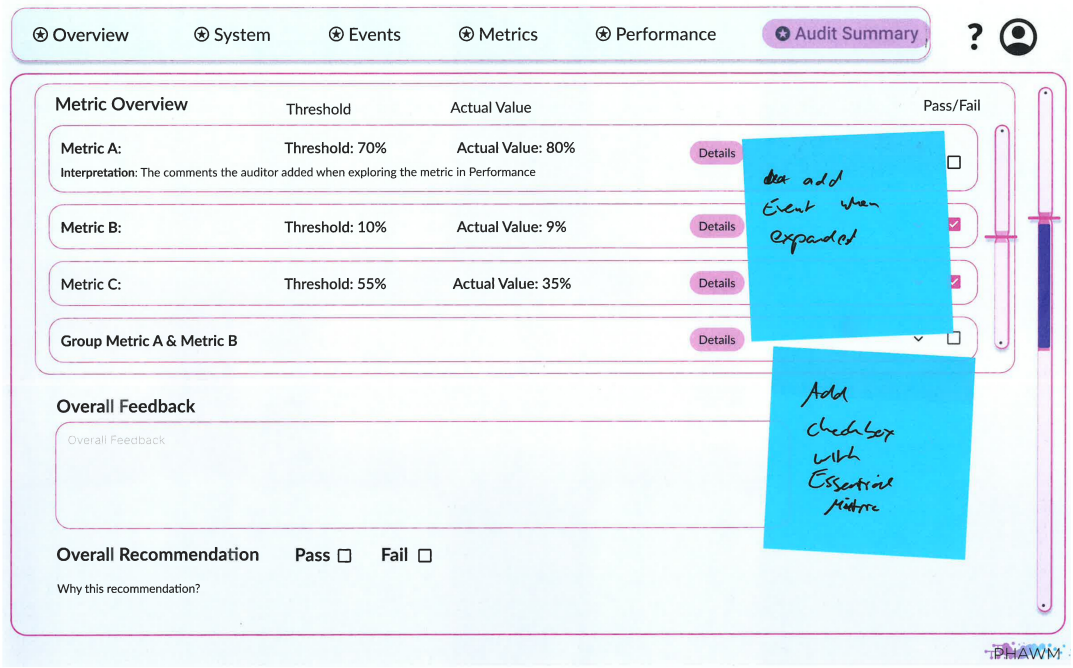


(b) Performance Analysis Provotype from Workshop 3 Parents.

Figure 18: Larger version of Performance Analysis Screens from Workshop 3 Patients and 3 Parents. The sketch shows a visual presentation of the metrics performance as line graphs, while the provotype focused on the scaffolding around graphs, which were indicated with empty placeholders. Blue post-its show feedback from the parents on the screen. Relating to Section 4.2.3 Figure 9.



(a) Audit Outcome sketch example from Workshop 3 Patients.



(b) Audit Outcome Provotype from Workshop 3 Parents.

Figure 19: Larger version of Audit Outcome Screens from Workshop 3 Patients and 3 Parents. The sketch shows a list of criteria with checkboxes, a textbox for notes and a clear pass/fail audit result option, which were all taken over into the provotype. Relating to Section 4.2.4 Figure 10.