“If it is easy to understand then it will have value”: Examining Perceptions of Explainable AI with Community Health Workers in Rural India

CHINASA T. OKOLO, Cornell University, United States
DHRUV AGARWAL, Cornell University, United States
NICOLA DELL, Cornell Tech, United States
ADITYA VASHISTHA, Cornell University, United States

Emerging work within the field of explainable AI (XAI) has shown promise for improving the interpretability of predictions produced by machine learning models but the benefits of such methods are often reserved for those with advanced technical knowledge of machine learning and AI. As the potential of AI begins to be realized in frontline healthcare, XAI can prove viable in explaining complex diagnoses, however, little work has been explored in this domain with community health workers (CHWs) who are the target users of many AI-driven tools in low-resource regions. We describe a qualitative study examining how CHWs engage with and perceive AI explanations and how might we design XAI interfaces that are more understandable to them. Through semi-structured interviews with CHWs and their deep engagement with a design probe to predict neonatal jaundice in which AI’s recommendation are accompanied by explanations, we (1) identify how CHWs interpreted the decision-making of the probe, (2) describe their perceptions of AI explanations, (3) unpack their reasoning for wanting explanations, and (4) detail how graphical and textual elements of the explanations impacted their understanding. Our findings demonstrate that while popular XAI approaches are not interpretable to CHWs, CHWs express a need for having explanations to improve communication regarding medical diagnoses and increase patient trust. We conclude by discussing what elements of AI need to be made explainable to end users, examining how better XAI can be leveraged through visualization methods, and highlighting potential design changes that could be implemented to improve the utility of XAI for socially and culturally diverse users with limited knowledge of AI in non-Western contexts.

CCS Concepts: • **Human-centered computing**; • **Computing methodologies** → **Artificial intelligence**; **Machine learning approaches**;

Additional Key Words and Phrases: Artificial Intelligence, Machine Learning, Community Health Workers, Mobile Health, Explainability, HCI4D, XAI4D, ICTD, Global South

ACM Reference Format:
Chinasa T. Okolo, Dhruv Agarwal, Nicola Dell, and Aditya Vashistha. 2023. “If it is easy to understand then it will have value”: Examining Perceptions of Explainable AI with Community Health Workers in Rural India. 1, 1 (April 2023), 32 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Authors’ addresses: Chinasa T. Okolo, Computer Science, Cornell University, 350 Gates Hall, Ithaca, New York, United States, chinasa@cs.cornell.edu; Dhruv Agarwal, Information Science, Cornell University, Ithaca, New York, United States, da399@cornell.edu; Nicola Dell, Information Science, Cornell Tech, New York, New York, United States, nixdell@cornell.edu; Aditya Vashistha, Information Science, Cornell University, Ithaca, New York, United States, adityav@cornell.edu.

XXXX-XXXX/2023/4-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Many low-income countries have established community health programs that have been shown to positively impact outcomes, including reducing neonatal mortality rates [12] and positively changing behavior [91]. These programs rely on the work of paraprofessional community health workers (CHWs) to provide last-mile health services [12, 144]. CHWs are usually women with high-school education who are recruited from local communities and receive a few weeks of medical training to provide essential health services based on approved health protocols [147]. They often work in challenging environments in hard-to-reach rural and urban communities. For example, CHWs deliver services like immunizations, family planning advice, and maternal-neonatal care in strongly patriarchal societies where as women they have limited agency. Their labor is often unacknowledged, undervalued, and unregulated [108]. They are severely underpaid, often receiving outcome-based remuneration, for example, US$8.40 for an institutional delivery.

Recognizing the structural barriers that CHWs experience, many researchers and practitioners have designed AI-driven tools to support the important work these low-skilled CHWs do. These AI-driven tools cooperatively work with CHWs to help them diagnose diseases [4, 14], analyze rapid diagnostic tests [38, 132], follow approved health protocols [5, 141], and manage patient-care [119, 130, 141]. Crucial for their collaboration is the ability of CHWs to understand and interpret the results of the models powering these AI-driven tools. This is especially important since prior work shows that CHWs often have very low levels of AI knowledge and place higher trust in AI's capabilities than their own [125], resulting in high-levels of overreliance [24]. In this environment, rushing to build AI-driven tools to support CHWs risks deploying AI in ways that might harm the very communities the CHWs aim to serve.

In response to the history of AI lacking interpretability and the resulting harm in high-stakes contexts like healthcare [14], a large body of scholarship in HCI and CSCW has emerged on making AI more explainable to developers as well as end users [2, 10, 28, 94, 99, 140]. However, most of the existing work on explainable AI (XAI) to date has focused on populations and contexts in the West. These findings may not adequately transfer to contexts like community healthcare in the Global South and socially and culturally diverse populations living in resource-constrained environments with low levels of digital skills and AI literacy [97, 168]. Although a nascent but growing body of work has started to explore how concepts of AI fairness and explainability differ across Western and non-Western contexts [124, 143], little is known about what end users like CHWs, with no or low AI literacy in high-stakes settings, need to know to become effective AI workers and cooperatively work with AI-driven tools. To fill this critical gap, we conducted a qualitative study to examine how CHWs in rural India engage with and perceive the effectiveness of current methods to make AI explainable. In particular, we sought to answer two research questions:

RQ1: How do CHWs engage with and perceive AI explanations?
RQ2: How might we design XAI interfaces that are more understandable to them?

HCI and CSCW researchers have frequently used technology provocations and design probes to better understand the needs of underserved communities and elicit elaborate responses through interaction with tangible objects [56, 75, 76, 169], especially in scenarios where target users lack digital skills and technology know-how [113, 125, 176]. Given low levels of AI knowledge among CHWs, we took inspiration from an existing AI-driven application for community health [37] to design a probe that cooperatively works with CHWs to diagnose neonatal jaundice. We implemented the probe in Figma and instrumented it to predict neonatal jaundice instead of using actual AI to make predictions. CHWs could use the probe to capture an image of a baby doll tinted yellow to simulate a jaundiced child, receive a prediction, and view explanations for how the application (probe) arrived at the prediction. We used two XAI methods, LIME [138] and SHAP [99], to show

explanations since these methods have been heavily used to explain computer vision models in diverse contexts in the Global South [124]. The explanations were designed to help situate the CHWs as users of XAI and to have them engage with and critically think about AI explanations in the context of disease diagnosis.

To answer our research questions, we observed 35 CHWs who interacted with the probe and conducted semi-structured interviews with them to examine how they engage with and perceive AI explanations. We also iteratively incorporated their feedback to examine how design changes to the current XAI interfaces might make them more understandable to the CHWs.

In response to RQ1, we found that the CHWs often questioned how the AI-driven application (i.e., our probe) came to a diagnosis. They compared the app to other diagnostic devices they used in their daily work that provide no explanations of the output (*e.g.,* thermometer, blood pressure monitors). Thus, they did not expect the app to explain how it arrived at a diagnosis. Moreover, they struggled to understand the notion of uncertainty in the app's diagnosis and viewed it as a definite decision rather than a prediction. Not only did CHWs blindly trust the app's diagnosis believing that it must always be correct, sometimes they even doubted their own expertise when their diagnosis differed from the app's prediction. Our participants were unable to understand the intended meanings of the SHAP and LIME visualizations, many of them explicitly stating that they were confused. They thought that the XAI interfaces depicted *symptoms* of jaundice rather than depicting the importance of the features that the AI used to arrive at a prediction. Surprisingly, despite the visible confusion in their interpretation of the XAI interfaces, the CHWs were vehemently in support of integrating explanations in the application stating that doing so would teach them new skills and improve the trust of their community in the app and in them.

For RQ2, we observed how the CHWs interpreted the LIME and SHAP visualizations, which rely heavily on color to highlight regions that contribute to the model's decision making. Surprisingly, the colors made it enormously difficult for the CHWs to understand the visualizations since they had strong preconceived notions of what the colors are expected to convey based on their prior experiences engaging with visualizations in the domain of Public Health. They also struggled to interpret the colorbars which explained the use of different colors in the visualizations. The visualizations also showed a reference image and other predicted classes, which further confused our participants. Other graphical and textual elements in the visualizations were also not interpretable.

Taken together, our findings show severe hurdles that could hamper effective cooperative work between low-skilled CHWs and AI-driven tools designed to support them. Based on our findings, we propose actionable design recommendations for future XAI visualizations that are understandable to end users with limited AI literacy and digital skills. Given the high levels of overreliance and AI technodeterminism, we discuss the need to design new XAI methods that encourage users to think critically and skeptically about AI outputs and scaffolding structures that enable novice users to meaningfully engage in cooperative work with AI-driven tools.

2 BACKGROUND AND RELATED WORK

We begin by providing background on community health workers (CHWs) and then describe the increasing integration of AI into their workflows. We then examine existing literature within the emerging area of research on explainable AI (XAI), detailing the XAI methods used within our study, describing research advances in human-centered XAI and designing XAI methods for non-technical users.

2.1 AI and Community Health Work

Many developing countries in the Global South rely on the work of paraprofessional community health workers (CHWs) to provide last-mile healthcare in low-resource communities [52, 63, 100, 118, 126]. CHWs are usually women with a high school education who are recruited from local communities, receive a few weeks of medical training, and then provide essential healthcare services to communities in hard-to-reach areas. CHWs provide a critical link between the community and the public health service, and are key to sustainable and resilient rural-urban health infrastructure [12, 147]. They have been shown to positively impact healthcare outcomes, including reducing neonatal mortality rates [12] and positively changing behavior [91].

Many HCI and CSCW researchers have taken a broad interest in developing tools to support CHWs in their daily work. These tools cooperatively work with CHWs to help them collect data to aid in monitoring and evaluating their patients [41, 65, 131], track supplies for distribution [38], receive feedback on their work performance [40, 42, 174], and improve their knowledge and skills [90, 112, 163, 164, 178]. These tools play an important role in supporting the work that the underserved, underpaid, and low-skilled CHWs do [77, 122, 136].

Recent advances in Artificial Intelligence (AI) have led to the widespread infusion of AI technologies that cooperatively work with health workers to address health inequities [20, 24, 35, 73, 86, 92, 98, 150]. Several HCI and CSCW researchers and practitioners have designed AI-driven tools that help CHWs in streamlining routine care [5, 141], disease diagnosis [37], scheduling visits [23], supervision of their tasks [130], and analyzing rapid diagnostic tests [38, 132]. More recent work in the Global South has examined CHWs' knowledge about AI and their perceptions about the use of AI in diagnosing neonatal pneumonia [125], incorporated AI into 3D anthropometry methods to aid CHWs to monitor the growth of newborns [5], used AI to analyze coughing sounds to potentially detect COVID in patients [4], aided nurses to detect diabetic retinopathy [14], and leveraged AI to improve maternal healthcare outcomes [119]. A growing number of scholars have also examined the promise and potential of AI in addressing last-mile healthcare problems [78, 104, 123, 128, 146, 167].

Despite the increasing integration of AI-driven tools into CHWs' workflows, little is known whether CHWs, who are primarily responsible for operating these tools in high-stakes settings, understand AI and what they need to know to become effective AI workers. This is particularly concerning since prior work shows that CHWs perceive AI to be infallible and lend substantial amounts of trust in these systems [125]. For effective collaborative and cooperative work to happen between CHWs and AI-driven tools, it is important to examine CHWs' understanding of AI and make its inner workings understandable to them. Our research attempts to address this critical gap by analyzing CHWs' interactions with and understanding of XAI methods, the potential benefits and challenges of integrating such tools into their work, and the potential impact on patient-XAI interactions. To motivate our research approach, we now look at prior work in explainable AI and the emergence of human-centered work within this domain.

2.2 Explainable AI

Explainable AI (XAI) consists of a set of methods that enable humans to understand the predictions made by machine learning models [10, 62]. XAI methods range from technical algorithms [80, 95, 99, 115, 138, 139], toolkits [57, 66, 74], and libraries [6, 18, 43, 53, 85, 121] to best practices and design principles [32, 54, 111, 116]. They help improve trust and allow users to understand the potential impacts of the underlying models specifically or AI systems broadly [47, 172]. Typically, XAI methods may be applied in local, cohort, or global ways [45, 114] to understand how the features within a model contribute to a single prediction, a set of predictions, or all predictions

produced by the model, respectively. The focus of our paper is on local explainability, as individual predictions are considered to be most relevant for end users of ML models [17].

LIME and SHAP are two popular methods to incorporate local explainability into models. LIME learns an explanation by testing variations of data in a machine learning model [138] and SHAP explains singular predictions by computing the contributions of each respective feature [99]. Our study employs the use of these two methods to explain predictions from a high-fidelity probe instrumented to diagnose neonatal jaundice. We use these methods to situate our participants as users of XAI and to encourage them to think critically about explanations. Both of these methods provide local explainability and leverage visualizations to explain image predictions produced from computer vision models. Our work contributes a novel empirical evaluation of LIME and SHAP, methods that have been frequently used to explain AI in diverse contexts in the Global South [124]. We now examine how researchers have began approaching the concept of centering humans in the design, development, and deployment of XAI methods.

Human-Centered XAI. The field of human-centered design advocates for researchers to engage with their target users before developing and deploying novel technologies [39], however, these practices are not commonly engaged within the development of XAI methods [124]. Ehsan et al. [49] introduce the concept of "Human-centered Explainable AI" (HCXAI) as a method that centers human users when designing XAI tools. Work in this space has focused on interviewing users and practitioners (UX designers, data scientists, researchers, etc.) to understand gaps in existing XAI tools [44, 48, 81, 96], evaluating XAI methods with various stakeholders [9, 25, 30, 46, 50, 84, 153], and introducing participatory co-design of XAI systems [111, 171, 177]. A growing body of scholarship has also begun to focus on how XAI tools support collaborative and cooperative work. Work critically examining the utility of XAI, especially in the scope of human-XAI collaboration has employed user studies that demonstrate a misuse of XAI tools [81], a lack of assistance provided to users in model evaluation [7], the potential of XAI to mislead users into accepting incorrect decisions [72, 81, 93, 134], and a limited ability for users to discriminate between correct and incorrect model predictions [11, 24, 84]. For example, Bansal et al. [11] showed how incorporating explanations into decision-making encourages human teammates to trust and accept recommendations from AI systems, regardless of their correctness. Kim et al. [84] also found *overreliance* [24] in their work evaluating visual explanations.

In high-stakes domains such as healthcare, this could lead to drastic medical outcomes for patients if healthcare practitioners overrely on AI-enabled tools to make decisions. However, there are ways to reduce this overreliance. Buçinca et al. [24] argued increasing people's cognitive motivation for engaging analytically with the explanations and developing effective explanation techniques to reduce overreliance. Vasconcelos et al. [162] showed that overreliance can be reduced by lowering the effort needed to understand and verify explanations, improving the efficacy of human-XAI collaboration [161]. Our study contributes to emerging literature in this domain by specifically focusing on CHWs in rural India who happen to be novice technology users and have limited experience interacting with AI. In particular, we examine how CHWs in rural India engage with explanations, how the explanations impact trust and reliance on AI, and how might we design explanations that are more understandable to them.

XAI for Non-Technical Users. HCI researchers strongly advocate against one-size-fits-all approaches [82, 168] and emphasize the need to cross-validate principles and measures with different populations [155]. As XAI methods continue to develop, it is increasingly important to empirically analyze XAI techniques with a wide range of stakeholders ranging from AI practitioners to non-technical end users. Existing research in this space has focused on understanding how non-technical users form mental models when interacting with explanations [33], proposed methods

to make models more explainable to non-technical users [15], and built a framework to support AI and design practitioners in creating XAI prototypes that cater to non-technical end users [79]. In relation to our study, several researchers have designed and evaluated visual-based XAI approaches to improve understanding for non-technical users [60, 117, 152, 156, 157]. For example, Shen et al. [148] examined the challenges the general public faces in understanding confusion matrices, a tool used to convey the performance of machine learning classifiers.

However, most of the existing work on XAI generally, and on non-technical users specifically, is focused on Western populations and contexts, which differ remarkably from the community healthcare context in the Global South. Additionally, there is little work focused on XAI in the Global South as evidenced in a review by Okolo et al. [124] which found that only a handful of studies have engaged with end users in the Global South to make AI explainable. Although a nascent, but growing, body of work has started to explore how concepts of AI fairness and explainability differ across Western and non-Western contexts [124, 143], little is known about what end users like CHWs, who have limited AI/digital literacy, need to know to cooperatively work with AI-driven tools in high-stakes settings. We contribute to this nascent domain by examining: **(1) How do CHWs engage with and perceive AI explanations in the context of AI-driven pediatric disease diagnosis?** and **(2) How do we design XAI interfaces that are more understandable to them?**

3 METHODOLOGY

To answer our research questions, we conducted a qualitative study in Summer 2022 with CHWs in rural India. To recruit CHWs, we partnered with a grassroots organization in Western Uttar Pradesh that runs several programs to strengthen community health systems in this region. A staff member from the organization contacted CHWs, explained the purpose of our study, and then scheduled interviews with the interested CHWs. All of our interactions with the organization and participating CHWs took place in-person.

3.1 Field Procedure

HCI and CSCW researchers frequently use technology provocations [76], exploration artifacts [113], and cultural probes [176] to inspire users to think about new technologies and better understand their needs in real-world settings. Given CHWs' low familiarity with AI generally and XAI tools specifically, we designed a probe to let CHWs engage with an AI-driven tool to detect neonatal jaundice in which AI's recommendations are accompanied by explanations. We incorporated popular XAI methods into the instrumented probe so that CHWs could engage with a tangible artifact and feel encouraged to think critically and concretely about explanations instead of abstractly. We implemented the probe in Figma and *instrumented* it to emulate AI's predictions and explanations.

Observation and Interviews.

We recruited 35 CHWs to interact with the probe and conducted semi-structured interviews with them. Three authors attended each session, with one leading the interview and the other two taking detailed notes and photos. The interviews took place in a closed room in community health centers to make it easier for CHWs to participate in the study. We first described the study to them and requested informed consent. After participants agreed, we then asked demographic questions to understand how long they had been working as a CHW, their smartphone usage, age, and education. We also asked participants about their knowledge of neonatal jaundice and how they diagnose it. We then described the functionality of the probe and used an iPad Pro to show the Figma prototype. We opted to use an iPad instead of a smartphone for its larger form factor. We provided life-size baby dolls and colored them yellow to simulate a real-life jaundiced child.

Fig. 1. A participant interacting with the probe.

Once a photo of a jaundiced doll was uploaded in the probe, it was *instrumented* to show: (1) diagnosis and severity level of jaundice, and (2) two XAI visualizations to explain the prediction. Our probe did not run a real machine learning model to make predictions. The Figma prototype outputted predefined predictions and associated visualizations. Before each interview, we decided which prediction to show to the next participant (e.g., moderate or severe jaundice).

After we showed a demo to the participants, we asked them to interact with the probe and closely observed them (see Figure 1). Once participants came to the diagnosis screen (before moving on to the explanations), we asked them many questions to examine their experience of interacting with the probe and their understanding of how the probe diagnosed jaundice. We then asked participants to think-aloud [27] as they interact with the first XAI visualization. Following this, we asked them several questions to gauge their understanding of the XAI visualization, including how this explanation helped them understand the diagnosis, and what they liked and disliked about the explanation. We also asked several follow-up questions about the specific explanation they had interacted with. For example, for the SHAP representation, we asked the participants what the colors of the squares, diagnosis text, and the colorbar meant. We then asked participants to interact with the second XAI visualization and repeated the process. At the end of their interaction with both XAI visualizations, we asked participants which explanation they liked better and why, and if they would find such explanations helpful if an app was developed to help them diagnose jaundice.

3.2 Probe Design

Representing Jaundice. We instrumented the probe to show moderate or severe levels of jaundice when a photo of jaundiced doll is uploaded to it. Jaundice is a condition that causes yellowing in the skin due to elevated bilirubin levels in the blood. It is a progressive disease that affects the lower regions of the body as it increases in severity. In newborns, jaundice is classified on five levels based on Kramer's rule (see Table 1) which illustrates the relationship between the progression of jaundice and its accompanying bilirubin levels [170]. To simplify these levels in our representations, we depicted explanations of moderate and severe jaundice in a manner in which we focus only on the regions important to a specific severity level. We thus represented explanations of severe jaundice as yellowness in the hands and feet, and moderate jaundice as yellowness in the shoulders, arms, torso, and legs.

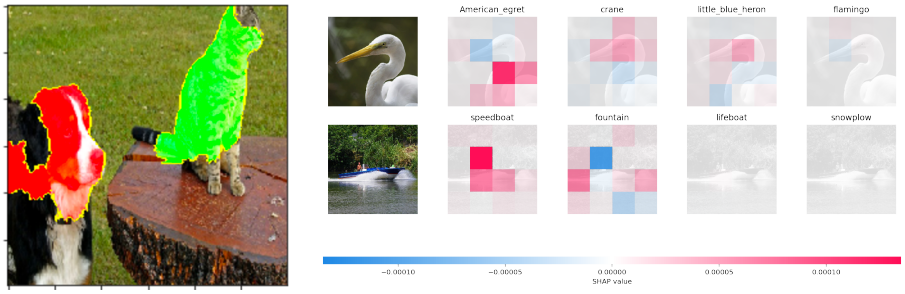
Area of the body	Level	Range of serum bilirubin	
		μmol/L	mg/dL
Head and neck	1	68–133	4–8
Upper trunk (above umbilicus)	2	85–204	5–12
Lower trunk and thighs (below umbilicus)	3	136–272	8–16
Arms and lower legs	4	187–306	11–18
Palms and soles	5	≥306	≥18

Table 1. Kramer's Rule for visual assessment of neonatal jaundice [170].

Explaining AI Predictions: LIME and SHAP. LIME [138] and SHAP [99] are the two most popular methods used by AI researchers and developers to incorporate local explainability into their models, especially when designing for underserved communities in the Global South [124]. We thus chose to use LIME and SHAP to explain hypothetical predictions from the high-fidelity probe designed to diagnose neonatal jaundice. Even though we suspected that CHWs might struggle to understand XAI methods given their low-levels of AI literacy, we incorporated these methods into the instrumented probe to elicit elaborate responses through interaction with a tangible artifact.

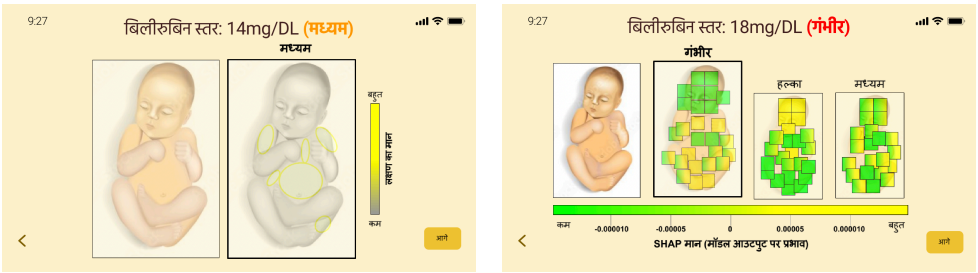
The default representation of a LIME explanation uses green to highlight pixels of an image that positively contribute to the predicted class and uses red to highlight the regions that negatively contribute to the predicted class (see Figure 2a). The default SHAP explanation produces a visualization depicting the original input image (as a reference image) with the images of the predicted class and other classes following it (see Figure 2b). On these images are squares that highlight how specific features (pixels) of an image contribute to the respective prediction. There is a color bar underneath these boxes with a gradient that ranges from blue (negative contribution) to red (positive contribution), indicating the respective SHAP value which is computed on running the SHAP library. Since the visualizations designed for our study were not produced from a trained ML model, there is no reference to what "actual" SHAP values would look like for our specific use case of diagnosing neonatal jaundice. So, we used the same scale as provided in reference examples found in the SHAP documentation (see Figure 2b).

Before conducting the study with CHWs, we organized a workshop with the staff of our partner organization to seek their feedback on the probe and the XAI visualizations. Based on their feedback, we made minor changes to the LIME and SHAP representations. The staff expressed concerns regarding the use of red color in the default representations of LIME and SHAP. They were worried that red color might confuse the participants since it is often used in public health messaging to imply "danger" in the medical sense. Since LIME traditionally uses red to highlight features with a negative contribution, non-jaundiced areas of the body would be red in the default representation, possibly indicating to our participants that such areas are contributing to the disease diagnosis. We thus changed the red (negative contribution) color to gray and the green (positive contribution) color to yellow, resulting in yellow for positive and gray for negative feature contribution in LIME (see Figure 3a). For SHAP, we changed the red color (positive contribution) to yellow and blue (negative contribution) color to green, resulting in yellow for positive and green for negative feature contribution (see Figure 3b). For example, to explain the prediction of severe jaundice, the SHAP visualization colored the hands and feet of the photo of the jaundiced doll in yellow, since these regions contribute more to a "severe" diagnosis. These color changes made our LIME and SHAP explanations slightly different from the original LIME and SHAP representations.



(a) LIME (b) SHAP

Fig. 2. Standard LIME and SHAP representations.



(a) LIME (b) SHAP

Fig. 3. Screens of the explanations used in our probes. At the top, the screens state the "bilirubin level" and the jaundice prediction: "mild" in (a) and "severe" in (b). In (a), yellow is used to highlight the bodily regions of the baby that positively contribute to the jaundice severity. The colorbar on the right serves as a legend for the "feature importance" from "low" to "high." In this case, since the shoulders, arms, torso, and legs are highlighted in yellow, the predicted class is "moderate." In (b), yellow boxes indicate pixels with higher SHAP values and green boxes indicate pixels with lower SHAP values, as explained by the colorbar labelled "SHAP value (impact on model output)." Higher SHAP values indicate a higher feature contribution to a respective class prediction. The three annotated images are labeled "severe", "mild", and "moderate" in decreasing order of the model's confidence. In this case, since the yellow boxes are primarily situated over the hands and feet, the predicted class is "severe", written in bold above the largest image. Since an ML model was not deployed in this study to output SHAP values, the numbers used on our plot are lifted from the standard SHAP representation in Figure 2b.

Gender	Female: 35, Male: 0
Age (years)	Min: 30, Max: 54, Mean: 42.8, Std: 6.6
CHW Experience (years)	Min: 4, Max: 17, Mean: 14.0, Std: 3.9
Technology Use	Computer: 0, Feature phone: 9, Smartphone: 26 (2 months – 10 years)
Education Level	Middle school: 13, High school: 15, Bachelors: 4, Masters: 3

Table 2. Demographic details of our participants.

Throughout our study with CHWs, we iteratively updated the XAI visualizations in response to the feedback we received from the participants to make the explanations more understandable to them. We describe these updates and their impact on CHWs' understanding in Section 4.

3.3 Participants

The CHW participants in our study lived and worked in rural Uttar Pradesh, one of the poorest states in northwestern India. They performed their work by traveling door-to-door, visiting patients at their homes, and providing them family planning advice, maternal and neonatal care, and essential health services based on approved health protocols. All of our participants were women, which is standard for practicing CHWs in India. They ranged in age from 30 to 54 years old and had an average of 14 years of experience working as a CHW. All of them owned or had access to a cellphone, with the vast majority (75%) using smartphones. Most of our participants had limited digital literacy and very low levels of AI knowledge and exposure. None of them used a laptop or computer. The majority of CHWs (80%) had at most a high-school education, with 13 CHWs completing middle school, 15 completing high school, 4 with a bachelor's degree, and 3 with a master's degree. We summarize the participant demographic information in Table 2.

3.4 Data Collection and Analysis

Throughout our fieldwork, we collected 15 hours of audio recordings and 57 pages of detailed notes. Audio recordings were translated into English and transcribed. We then used inductive thematic analysis [21] that allows key themes to emerge from the raw data through repeated examination and comparison. The two co-first authors led the qualitative coding process. To begin, we each coded three interviews separately and met to reconcile conflicts by merging similar codes and organizing the remaining codes into a streamlined codebook. We continued to code separately and came together after each interview until we had reached intercoder reliability and codebook stabilization. Following this, we separately coded the remaining transcripts. Throughout the analysis, we held multiple discussions to iteratively refine the codes and used peer-debriefing to reconcile disagreements. After conducting multiple passes, we ended up with 121 codes grouped into five themes: General XAI Interpretations, Probe Interpretations, XAI Preferences and Suggestions, SHAP, and LIME. For example, the "General XAI Interpretations" theme categorized CHW interpretations of features that were consistent across all explanations like bar graphs, colorbars, and images. Examples of such codes include: "XAI Interpretation: bar graph", "XAI Interpretation: gradients", and "XAI Interpretation: reference image."

3.5 Ethical Considerations

We gained IRB approval for all study procedures. Given the COVID-19 pandemic, we were vigilant in protecting the safety of the study participants. Masks and hand sanitizer were available and study implements (iPad and baby dolls) were sanitized regularly. We were also flexible in scheduling the interviews since CHWs had to juggle their schedule to provide patient care. Participants were remunerated in kind by our partner organization for their involvement in the study.

4 FINDINGS

All CHWs struggled to understand AI explanations. While this was expected, through a deep engagement with the probe, they provided elaborate responses on how they perceive the benefits and pitfalls of AI explanations and how different features in XAI visualizations impact their understanding. We begin this section exploring our first research question to understand how CHWs interpreted AI (Section 4.1), how they perceived explanations in the context of their work (Section 4.2), and their reasoning behind wanting explanations (Section 4.3). For our second research question, we then discuss how different features of LIME and SHAP visualizations like color (Section 4.4), graphical elements such as images and shapes (Section 4.5), and textual elements (Section 4.6) impacted CHWs' interpretation of the explanations.

4.1 How CHWs Interpreted AI

At the beginning of the interviews, we asked participants if they were familiar with AI. Most CHWs had not heard of AI and did not know what it is. After we showed the participants the app (probe), we asked them to describe how they think the app diagnoses jaundice in newborns. While almost all CHWs understood what the app does (e.g., help diagnose jaundice in babies), they struggled to pinpoint how the app works. In particular, they were unable to *explain* how the app arrived at the prediction, often being clueless and comparing the process to *"magic."* In their everyday routine, they relied on several heuristics to determine if a baby has jaundice, including looking at the color of the skin and temperament of the baby. However, only a few CHWs were able to guess that the app must be looking at the skin color to detect yellowness, particularly since they had no exposure to end-user applications that use computer vision.

Since the CHWs were unable to understand how the app arrived at a prediction, they relied on their own knowledge of jaundice and neonatal care to interpret the diagnosis and accompanying explanations. CHWs would often look at the doll on the table and use their knowledge of expected "symptoms" to make sense of the prediction: *"The baby's feet and stomach looks yellow, that's why the app has encircled that region."* (P35). They often used their perception of how the baby (doll) is "feeling" as a key factor in their respective decision-making, stating things such as *"The doll looks cranky, that's why the app has detected jaundice."* (P18), something which is very difficult for AI to do in practice [13, 120]. In a way, the CHWs were ascribing superior capabilities to the AI-driven app than it actually possessed.

Comparison to Diagnostic Devices. CHWs' prior experience with medical diagnostic devices shaped their AI-related mental models. Throughout the interviews, CHWs often relied on their experiences working with diagnostic devices to arrive at an understanding of how the app might be detecting jaundice. Participant P02 compared the probe to blood pressure monitors and thermometers and stated, *"How can I tell how these machines work? They just do."* Similarly, P04 stated, *"The same way a thermometer checks for fever, this app is checking for jaundice."* CHWs stated how in most of the medical equipment and diagnostic tools they used in their daily work, no explanations are provided in the output and they also do not know how such devices work under the hood. For them, a thermometer simply provides a temperature, a scale provides a weight, and a COVID-testing system gives a negative, positive, or sometimes an inconclusive diagnosis. Additionally, CHWs stated that detailed reports from medical tests are commonly interpreted by a trained doctor, not by them. Hence, the concept of needing to understand exactly how the probe comes to a specific jaundice diagnosis was quite discomforting to our participants. However, some CHWs conceded that they need to know *enough* to operate these devices safely and understand the output to a degree where they could explain it to the community members. "What" these diagnostic tools do mattered more to them, instead of "How" they do it. These findings show that when CHWs interact with AI-enabled tools designed to augment their work, they are prone to rely on their prior experience operating medical devices and providing patient care to discern how such tools operate.

Discomfort with Uncertainty in a Machine's Output. We also observed that the CHWs often treated the output of the probe as an absolute decision instead of an estimation or prediction. They assigned the same level of trust to the AI app as they would to any diagnostic system like a thermometer, blood pressure monitor, or weighing scale, believing that the app could never be wrong. For example, P08 emphasized, *"a machine can never be wrong"* and P15 felt that *"since the app is a computerized system, it is natural that it would give the right result."* Moreover, when the explanations showed them things that went against their intuition, instead of questioning the prediction

and the underlying AI, they started doubting their own knowledge and understanding of jaundice. They believed that they were the ones uncertain about their diagnosis and the machine would help them resolve their doubts by telling the truth. Past work has also uncovered overreliance on AI, *e.g.*, accepting incorrect decisions without verifying whether the AI is correct [24]. However, this problem becomes more severe when low-skilled CHWs use nebulous AI systems in high-stakes settings where results must be interpreted with caution, considering that they did not understand confidence values or probabilities in context of AI predictions.

4.2 How CHWs Perceived XAI Explanations

In addition to understanding participants' perceptions around the diagnostic capabilities of the app, we also explored their reactions to the provided explanations. Most CHWs struggled to understand the explanations of how the underlying AI arrived at decisions. Although a few participants claimed to understand the explanations presented to them, the understanding they voiced was usually far from the intended meaning and their interpretations kept shifting. For example, P17 said the two images in LIME were of the same baby, but later said that the baby in the second image had more severe jaundice than the baby in the first image, contradicting her claim that they were photos of the same baby. Other CHWs thought that the SHAP explanation showed a story of a baby transitioning from having no jaundice to severe jaundice. While these interpretations made sense to the CHWs, they were not what the XAI method intended to convey. Only a few CHWs were able to partially understand the explanations with the help of higher-level features such as colors, images, or shapes. For example, P07 understood that the SHAP explanation was showing varying severities of jaundice: *"In the severe image, the entire body has yellowness. In the mild one, it's only the face. In the moderate case, the entire body has yellowness but it is light yellow."*. Almost all CHWs explicitly stated their confusion, often asking the interviewer to help them understand the explanations. P06 expressed the importance of understanding explanations in relaying the app's prediction to her patients:

> *"I can't understand these explanations. If I can't understand them, then how can I take the app to my community and explain them what this [explanations] is? I have no confidence that I understood it properly."* (P06)

Most CHWs already placed high trust in the diagnosis by AI and the presence of explanations further reinforced their trust in the app. These views of our participants are concerning because existing research has shown the possibility of explanations to reinforce incorrect predictions [84], which in the case of healthcare could have severe consequences.

Symptoms vs. Feature Importance. The CHWs construed the XAI designs to depict the *symptoms* of jaundice in the baby, rather than depicting important contributors to the app's prediction. To them, the visualizations showed parts of the baby that have jaundice, not parts of the baby that led to the prediction. For example, P05 noticed the outlines on the LIME representation and said that the line represents parts of the baby that are infected with jaundice. In SHAP, P07 interpreted the yellow boxes as body parts infected with jaundice and green boxes as healthy body parts. In reality, while some body parts may appear more yellow than others, jaundice does not affect different body parts with different severity. The entire body is said to have jaundice, not isolated body parts like the hands or feet. When we mentioned this to the CHWs, they said they knew this fact. However, their interpretation remained that some parts of the body had "more" jaundice than others because of their constant confusion about the feature importance (i.e., body part is yellow to indicate the importance of the feature in decision-making) versus the symptoms of jaundice (i.e., body part is yellow due to jaundice). For example, P35 got confused when we asked her why the baby's foot is not yellow. She replied, *"There is less jaundice in the foot, more*

in the rest of the body." Like most other CHWs, she alluded to yellowness in the body, which is a mere symptom of jaundice. In reality, the lack of yellowness in the foot indicated to the model that the baby might have moderate jaundice and not severe jaundice since yellowness in the foot is a symptom of severe jaundice as per Karmer's rule (see 1).

4.3 Why CHWs Still Wanted Explanations

Although CHWs struggled to understand the explanations, they still expressed a strong desire to have explanations when cooperatively working with AI-driven tools. When articulating the importance of having explanations, CHWs mentioned various reasons such as, explanations improving their own comprehension of the diagnosis, improved community trust, and usefulness in explaining to others how the app works.

Better Understanding of Diagnosis. When we dug deeper to understand why CHWs preferred to have explanations even though they struggled to understand them, they mentioned that the explanations would allow them to understand how the app arrived at a diagnosis, especially once the explanations are made easy to understand or CHWs are trained to interpret them. P09 emphasized the utility of the explanations in making her understand the diagnosis, but warned against their interpretability,

> *"See, they both [LIME and SHAP] are very confusing. I think you should take it out of the app. Keep it only if we can understand. If it is there and is easy to understand then it will have value. We will have more information and we will not forget things."*

This shows that explanations are useful to CHWs only if they can be interpreted. Many CHWs believed that an AI-driven tool in which AI's recommendations are accompanied by explanations would serve as a tutor, teaching them new skills and providing them hands-on training to detect jaundice. However, a few CHWs were also critical of the explanations. For instance, P02 wanted only a decision from the app, not an explanation. She saw value in using explanations to tell the baby's parents how the app came to a diagnosis, but did not feel that it helped her understand the diagnosis better.

Improved Community Trust. Our participants often emphasized the tireless efforts they do to build trusting relationships with community members, especially since they work in communities with strong patriarchal systems where, as women, they have limited agency. It was of great importance to our participants to be able to relay detailed information about a diagnosis to the broader community. Since no such applications are currently used in the community, CHWs felt that the explanations would demonstrate that the app does not produce decisions in an arbitrary manner and increase the community's trust in the app and in them. P03 highlighted how the explanations would prove that the app did not arrive at a decision randomly and engender trust in them:

> *"People will understand how the decision is taken easily if the app is there...and that they need to see the doctor ASAP."* (P03).

P04 and P05 emphasized that their patients trust them more when they use visual aids. They gave the example of paper-based visual aids (see Figure 4) they carry to educate pregnant women about when to seek urgent care. When we asked if CHWs would prefer similar paper-based visual aids for jaundice, P04 and P05 preferred the AI-driven app over paper-based visual aids, stating that the app could diagnose the severity of diseases which the paper-based aids could not. P06 shared:

> *"The app is useful because it tells how much the baby has jaundice. I can't do that. The public would also trust the app more...If we use the app, we will also learn how the app*

Fig. 4. An example paper-based visual aids that CHWs used to improve community trust in them.

> *diagnoses jaundice, which will also help us improve and share the information with community members."* (P06)

CHWs mentioned that the explanations would also allow parents to take a diagnosis more seriously. As P08 put it, *"they will understand why their baby has jaundice and what are the associated symptoms."* While this was perceived to be a helpful outcome of including explanations, some participants debated the right amount of details and the language to be included in the explanations to not cause panic. P12 argued, *"explaining too much would cause the villagers to worry or panic."*

We now describe how different XAI design elements (e.g., color, graphical elements, shapes, and textual elements) impacted participants' understanding of AI. Although these findings are in the context of LIME and SHAP visualizations, they also broadly apply to other XAI methods.

4.4 How Color Impacted CHWs' XAI Understanding

Color was used in both the graphical (shape and image) and textual elements of our LIME and SHAP explanations. In this section, we detail the usage of color within LIME and SHAP and how it impacted the CHWs' understanding of XAI visualizations.

Colorbars. Colorbars were used in the XAI visualizations to indicate the contribution of an underlying feature to the model's decision-making. In LIME, the colorbar was situated vertically with a color gradient indicating a low feature value at the bottom and a high feature value at the top (see Figure 3a). In SHAP, the colorbar was situated horizontally with a color gradient indicating a low feature value at the left and a high feature value at the right (see Figure 3b). Only one out of 35 CHWs was able to use the colorbar effectively to form an understanding of the XAI visualizations. The rest either did not notice the colorbars or struggled to interpret them. They were unable to map the colors on the colorbars with the colors used within the visualizations. While some participants (e.g., P07, P17, and P18) noticed that there was more yellowness at the top and less at the bottom of the LIME colorbar, they were unable to relate this information to form an understanding of the LIME visualization. P17 perceived this difference in color to mean that the baby would still have some chance of surviving unless the colorbar became fully gray, while P04 incorrectly interpreted the colorbar as a proportional scale of severity of the disease:

> *"There is more yellowness at the top and lesser at the bottom... Three-fourths of this bar is yellow and the remaining is gray... I think this means that out of four parts, the kid has jaundice in three parts."* (P04)

While CHWs noticed the text accompanying the colorbars, it did not improve their understanding. For example, P17 incorrectly interpreted the meaning of "less" to mean that there is a lack of blood and "more" to mean that there is more yellowness in those areas. These findings show that the CHWs in our study found the colorbars confusing and often misinterpreted the conveyed meaning, suggesting the need to design simpler visualizations for novice users.

Mental Models of Color. CHWs often relied on their prior experience with colors in visual aids in the domain of Public Health to interpret what the colors in the XAI visualizations might mean. CHWs had strong color associations, believing that *"green means safe, yellow means warning, red means danger, and gray/blue means lack of blood."* These color associations largely influenced how they interpreted the visualizations depicting a jaundiced baby.

We see an example of this in P06's interpretation of the LIME visualization, which highlighted in color the parts of the body that played a role in the prediction of jaundice (see Figure 3a). She felt that the baby was in moderate danger because the body had patches of yellow color. When we asked why the face of the baby is gray, P06 felt that this is because the baby has a darker skin color. This interpretation was incorrect because the baby's true skin color was only visible in the reference image and the explained image was overlaid by the colors of the visualization (yellow or gray in this case). In addition to showing the impact of mental models of color on interpretation, this example suggests that CHWs struggle to understand what colors meant, especially given the failure of colorbars to explain the meaning of different colors.

Some participants also noticed that the outline of shapes was darker than the fill color of the shapes. For example, the fill color was light yellow and the outline color was a darker yellow, to mark it off from the rest of the body. P20 noticed this subtle difference and interpreted that, *"since the outlines are darker, jaundice is more severe in the outer areas and it is trying to spread more to the rest of the body."* This shows that not just the choice of color but also different shades of the same colors may be interpreted differently, suggesting that the color of even subtle XAI features like shape outlines should be carefully chosen and clearly explained.

CHWs' context also shaped their interpretation of the SHAP visualization, which showed boxes colored on a green-to-yellow gradient (see Figure 3b). Yellow boxes depicted pixels that contribute positively to the model output, green boxes showed pixels that contribute negatively, and half-green half-yellow (henceforth "green-yellow") boxes depicted moderate contribution to the output. Since the CHWs could not understand the accompanying colorbar, they came up with their own reasoning about what the boxes might mean. Based on their prior color associations, they assumed the presence of jaundice in the body parts containing yellow boxes and the absence of jaundice in body parts containing green boxes, but these boxes indicated contribution to the model's decision-making instead of the presence of jaundice in different body parts. Moreover, the green-yellow boxes confused them even more.

Given the CHWs' strong associations with colors like red, green, yellow, blue and gray, we explored how a color scheme that did not include these colors would impact CHWs' understanding of the XAI visualizations. Through discussions with CHWs and the staff of partner organization, we designed a new LIME visualization (see Figure 5a) that used a pink color scheme: dark pink to represent positive and light pink to represent negative impact on the prediction. Interestingly, even though pink seems close to red, it did not evoke a feeling of 'danger'. Instead, participants assumed that the dark pink color suggests higher jaundice in the highlighted body parts, a partially accurate assumption however not something which the visualization meant to convey. Given that the CHWs had weaker existing mental models of pink being used in the context of disease diagnosis in their daily work, their assumptions were more measured. These findings show that CHWs have strong mental models of what the colors are expected to convey in visualizations in the community

(a) LIME (b) SHAP

Fig. 5. Additional re-designed LIME and SHAP explanations. In (a), some body parts are encircled in a dark pink color and the rest of the body is shaded light pink. On the right is a descriptive legend saying, "Which body parts contributed to the prediction of moderate jaundice?: [Dark Pink]: More contribution, [Light Pink]: Less contribution." In (b), the three images show the three severity classes, labeled "moderate", "mild", and "severe" jaundice. The current predicted class is in bold ("moderate") and of the largest size, with each following image smaller in size, depicting decreasing confidence in those classes. On each image are green and yellow squares. The legend on the right contains exactly the same text as in (a), with green squares for less contribution and yellow ones for more.

(a) LIME (b) SHAP

Fig. 6. Re-designed LIME and SHAP explanations without colors. In (a), the circular outlines lead to a descriptive text saying, "These body parts contributed more to the prediction." The text box on the bottom says, "Rest of the body contributed less to the prediction." In (b), the three images show the three severity classes, labeled "moderate", "mild", and "severe." The current predicted class is in bold ("moderate"). Above each image is a confidence label, "most", "a little less", and "least" confidence. On each image are shapes: squares to represent less contribution to the prediction, and circles for higher contribution, as is explained by the descriptive legend on the right.

health context. In the absence of any training to interpret visualizations and read colorbars, they relied on their prior experience to assume what the visualizations meant. This suggests that XAI visualizations in the community health context need to: (1) carefully choose colors that have no strong prior associations, and (2) clearly describe what the colors mean.

Prototypes Without Color. Due to the ongoing confusion with the meaning of color and the failure of colorbars to aid CHWs in interpreting the feature importance, we later adjusted the probe to not use any colors.

In SHAP, we replaced colored boxes with shapes: circles for higher feature importance and squares for lower. To accommodate this change, we removed the colorbar and instead added a

discrete legend with a descriptive title (seeFigure 6b). Our main aim was to design XAI interpretations that were not affected by CHWs' mental models of color. We did not want them to interpret one color as "more jaundice" and another color as "less jaundice" in the body parts, but instead as "more" or "less" contribution of the body part in the AI's prediction (i.e., *feature importance*). Upon using these shapes, CHWs started to move away from the jaundice/no-jaundice interpretation that was present when colored boxes were used within the SHAP visualization. Since they did not have colors to base their explanations upon, the participants started to use their understanding of jaundice to interpret the visualization. For example, P29 pointed to the circles and said that in her experience, the encircled areas most commonly help with diagnosing jaundice. Further, since the shape did not have a color fill, CHWs could now notice that the baby's encircled body parts looked yellow and use this information to guess the severity of the disease. With some hand-holding, CHWs that were shown this probe were able to get an initial understanding of the correct interpretation of SHAP. This shows that introducing colorless shapes along with descriptive legends evoked richer responses and better understanding since the CHWs were no longer limited to their mental model of colors. Additionally, the CHWs were quick to identify the two different shapes used in this probe. This also highlights how the discreteness of these shapes enhanced their ability to distinguish between the parts of the body that did and did not contribute to a respective prediction.

In LIME as well, we designed probes that tried to depict the same information but without using colors. As seen in Figure 6a, in this version, we outlined the parts of the body with less and more contribution and labelled them with text descriptions. While there was some inertia in the CHWs to read large blobs of text, once they did, some of them started to form an initial understanding of the explanation. Devoid of color, this version made it easier for some CHWs (P25, P33, P35) to move towards a more correct understanding of what LIME intended to explain. P33 said:

> *"The stomach, feet, chest, shoulders [encircled body parts] help in identifying moderate jaundice. These body parts are playing a larger role in telling that the baby has jaundice. The rest of the body parts are informing the computer less about jaundice."* (P33)

However, there was still confusion about the specifics. While the size of a respective circle was not indicative of feature importance, one participant mentioned, *"The stomach has the most contribution because a bigger circle is drawn over it."*(P31). In reality, the stomach only had a bigger circle because it was a bigger region.

These findings show that removing color from the visualizations was helpful for some CHWs. It prevented them from relying on their existing mental models of color associations, thereby improving their understanding of how the probe estimated jaundice severity. However, only a few CHWs benefited from this change. Most participants continued to view the encircled body parts as symptoms of jaundice rather than the explanations of the prediction. We conclude that XAI visualizations without color are more interpretable than visualizations involving color, but are still not fully understandable for most CHWs.

4.5 How Graphical Elements Impacted CHWs' XAI Understanding

The LIME visualization in the app displayed two images of the baby (see Figure 3a): the input image and an annotated image describing feature importance. The SHAP visualization displayed four images of the baby (see Figure 3b): the input image, and an annotated image for each severity class (mild, moderate, and severe). The input image is unlabelled and serves as a reference image, and the three annotated images represent the model's evidence for the three severity classes in a

decreasing order of AI's confidence. Multiple images of the baby in the same visualization caused a lot of confusion for our participants.

For example, CHWs were often confused by the input image. They were generally unable to explain its presence. In our subsequent prototypes, we wrote the literal text "The image you entered" in Hindi on top of the reference image, but even then most CHWs were unable to understand that this image was just for reference. Instead, they provided varying interpretations, such as the image sequence is demonstrating the progression of jaundice (P03, P06), they represent multiple babies with different levels of jaundice (P04, P05), or the baby in the reference image being *"less"* sick than the ones in other images (P06, P07, P23). P06 stated:

> *"The baby in the reference image looks weak and inactive. But this baby has no jaundice because the hands and legs are relatively less yellow."* (P06)

In the SHAP visualization, the CHWs interpreted the labels for the other three images literally and assumed that the baby labelled 'mild' has mild jaundice and the one labelled 'severe' has severe jaundice. However, these severity labels meant to represent explanations for other classes had the model actually predicted those severity classes. Further, some CHWs tried to correlate the severity label to the number of green and yellow boxes in the respective image. In case of the severe jaundice prediction, the labels confused CHWs and they assumed that the baby labelled severe has the maximum jaundice because it has the most number of yellow boxes on it. This interpretation is incorrect because the underlying "model" relied on specific regions of the body (in alignment with the Kramer's Rule chart in Table 1) to provide a severity prediction. For example, a "severe" prediction would primarily have yellow boxes covering the palms and soles of the feet compared to a "moderate" prediction that would have yellow boxes covering the shoulders, arms, torso, and legs.

In case of the moderate jaundice prediction, the image labelled moderate had more number of yellow boxes than the image labelled severe. This is because the yellow and green boxes do not represent the severity of jaundice in the baby, but the importance of different pixels in the prediction. Some CHWs noticed this themselves and pointed it out to us, either suggesting that the images are labelled wrong or that their understanding is flawed. Other CHWs did not notice this disparity, and when we pointed it out to them, were left confused about their understanding.

Some CHWs understood that the four images represent the same baby, but incorrectly assumed the order of these images to indicate disease progression. For example, P04 and P05 felt that the baby had a mild infection initially but had the risk of getting moderate or severe infection if left untreated. Similarly, P28 assumed that the images represented a chronological progression of jaundice: first, the baby was sick (moderate label), it was given medicine to get better (mild label), but then relapsed into severe jaundice (severe label).

Only a few CHWs noticed the number line below the images in the SHAP visualization. For example, P02 observed the numbers on the line increasing from left to right, but assumed that those numbers represent the severity of jaundice. Instead, the numbers indicated a quantitative (negative, positive, or neutral) feature contribution of the green and yellow boxes. P02 also interpreted the placement of the images as being labeled in increasing order of severity. However, the images were not labeled in increasing order of severity, but in increasing order of the model's confidence in the predicted severity (e.g., sometimes 'severe' was displayed before 'mild'). To justify this inconsistency between the probe image ordering and how they thought the images should be ordered, some of the CHWs stated that the *"images were displayed in an incorrect order."*

4.6 How Textual Elements Impacted CHWs' XAI Understanding

Both LIME and SHAP visualizations had a text placed at the top of the screen to indicate the app's prediction. CHWs almost always read this text, but did not pay attention to other textual elements. Instead, they primarily focused on the graphical elements on the screen. For example, in our final LIME prototype, we removed all colors and the reference image, and had just one image with an accompanying text to explain the shapes drawn on that image (see Figure 6a). Even then CHWs did not read the text without being asked to and only focused on the image. We had to often ask CHWs to read the text accompanying colorbars, legends, etc. and describe what the text meant. Even though over 60% our participants had at least a high-school education, almost all of them were unable to understand the mathematical textual elements. They did not notice the minus sign and the decimals used in the number line in SHAP, for example, many CHWs read -0.000010 as just 10.

Confidence Values in SHAP. As described earlier, the images representing different severity classes were presented in a decreasing order of AI's confidence, which greatly confused our participants. In the iterated versions of the probe, we added explicit confidence labels to the images: as percentages in one probe (e.g., '95% confidence') and as descriptive text in another probe (e.g., 'least confidence'). However, these labels also did not improve the CHWs' understanding of the confidence intervals. For example, P28 ignored the % sign and interpreted "95" as body temperature since body temperature is usually measured in Fahrenheit in India and 95° is slightly below normal. However, she soon got confused as 3° (for '3% confidence') and 2° (for '2% confidence') did not make sense as body temperatures. P24 tried to interpret the confidence percentages slightly differently, but got confused along the way:

> *"These numbers show decreasing symptoms of jaundice: 95, then 3, then, 2. But, why is the moderate image listed under 95 and the severe under 3? I don't understand."* (P24)

In addition to changing the severity labels in one probe, we also explored changing the sizes of the three images such that the decreasing size of an image represents decreasing confidence. For example, the most confident prediction shown is the largest image and the least confident prediction is the smallest image (see Figure 5)b). Most CHWs did not even notice the different sizes of the images. When we asked them to, the varying sizes did not help them in interpreting the confidence values. Many CHWs (e.g., P20, P21, P22, P24, P25) assumed that these are children of different ages, depicted here to show that jaundice can happen to kids of any age. P28 was utterly confused:

> *"If the first image was the smallest, then I would understand that the baby is growing in size with age, but here the first baby is the biggest... Oh, maybe the baby is getting drier due to the illness [meaning that its weight is reducing]?".*

As we emphasized earlier, the meaning of confidence itself was unclear to our participants in this context. When asked to guess the meaning, P23 said that *"these labels show what people in the field should trust more,"* which is true as an outcome of the confidence values, but not their explicit meaning in the XAI visualization. Multiple CHWs (P24, P25, P29) interpreted the confidence levels depicted in the probe as reflecting *their* own respective confidence. For example, P24 said, "The app is showing how much confidence I should have in saying that the baby has moderate jaundice." These interpretations also shaped our choice to change the label in the app from *decision* produced from the probe to a *prediction* to enforce the mental model that AI is merely *predicting* a jaundice severity and not providing a firm diagnosis. Our findings suggest that such actions might be necessary to help CHWs retain their agency when acting on predictions produced by ML models in high-stakes situations.

5 DISCUSSION

5.1 What elements of AI need to be made explainable to end users?

In our study, some of our participants questioned the need for XAI stating how they only found the decision necessary. However, many of our participants expressed opposing views, stating how XAI would not only be helpful for them to improve their understanding of disease diagnosis but would also aid in guiding their interactions with patients when they may be tasked with explaining how the system came to a decision. Our participants were able to see these benefits despite being unable to interpret the LIME and SHAP representations shown to them. However, they still expressed a strong need for explanations that are actually interpretable. In line with many of our findings, the ability of XAI to complement human work still shows mixed results, with many of the benefits touted by XAI often reserved for the developers of these models [17, 31, 44, 89, 180]. Additionally, researchers have shown that these methods can provide conflicting explanations [88], encode trust in incorrect decisions [11, 84], and increase model complexity [74]. To work towards improving the utility of XAI, AI practitioners and researchers will need to be aware of these constraints and actively work towards addressing them.

The concept of AI explainability may improve if we approach XAI methods by examining how to help users think more *critically* about AI in general. In contexts like community health, we run the risk of users ascribing more qualities to AI than it holds and the potential for over-reliance on AI predictions [125]. There is a current need for methods to keep such ideologies in check and XAI can potentially keep users critically questioning how ML models operate and prevent errors from harming vulnerable populations. Current explanation methods do not give end users any level of agency to rebut incorrect predictions, something that researchers and legal frameworks such as The EU General Data Protection Regulation (GDPR) frame as a right for users of AI systems [36, 137]. Given the context of our work with CHWs, new interpretability methods that give workers reason to question, engage, and think critically about AI can help prevent users from forming false beliefs about explained ML models [33] and provide these users with the agency to actively challenge these systems.

Our experiences in the study also highlighted potential design changes that could be implemented to reduce our participants' confusion and visible frustration with the explanations. Springer et al. [154] conducted two studies evaluating how users react to transparency in systems to build upon the concept of "progressive disclosure", originally introduced by Xerox to ease users in operating systems new to them by hiding complicated features of the user interface [26]. Their work suggests that extremely detailed transparency information (factors influencing the output of ML models) may not always be important to users and provides a novel design approach to improve transparency of algorithmic systems by providing users with the autonomy to view explanatory features to their respective preferences. By prioritizing this concept, we can prevent cognitive overload by showing specific features of explanations in an ad-hoc manner, where end users are shown explanations to their respective preferred level of detail. We believe that explanations can also cater to the needs of end users by combining progressive disclosure with the concept of "personalized explanations", a method introduced by Schneider et al. [145] to customize explanations to users based on *what* information should be depicted, *how* to depict this information, and *how much* information should be included.

There is a need for more research focused on centering end-user needs while also accounting for an emerging set of users with low digital skills and no AI experience. So far, there exists limited work specifically focused on developing end-user explanation methods [19, 33, 79]. Work by Jin et al. [79] identifies user-friendly forms of XAI to propose the "End-User-Centered Explainable AI Framework" which taxonomizes these methods into 12 categories and provides a way for

researchers to use the categories to build novel, user-friendly XAI interfaces by leveraging prototyping and participatory design. While findings from this work may generalize to "non-technical" populations, more work is needed to examine how these findings apply to socially and culturally diverse users in the Global South, such as our low-skilled CHW participants who not only have low digital/AI literacy but also have limited domain expertise in advanced medical diagnoses and procedures. To bridge this gap, researchers need to work closely with such users to understand their segmented explanation needs and build new methods to actively address these needs.

5.2 Improving Access to XAI in the Global South

Our work shows that when providing explanations in the community health context, there exist many intricacies when deciding *how* and *what* to explain from a model result and also *who* will benefit from the explanation. Researchers have already begun work in this area, providing an overview of challenges within existing XAI methods [107, 110, 124], examining the types of stakeholders that need explanations [16, 22, 44, 69, 94, 135, 158], guiding practitioners to develop effective human-in-the-loop AI systems [8, 58, 109], and developing measures to evaluate the effectiveness of the explanations [3, 70, 71, 84]. While this literature provides strong depth into understanding how AI researchers and practitioners can evaluate the efficacy of XAI and design more user-centered XAI methods, there is much room to improve AI understanding of socially and culturally diverse users in non-Western contexts who have low digital/AI literacy and are expected to cooperatively work with AI tools in high-stakes settings.

To understand what such socially and culturally diverse users need to know to be effective AI workers, what technical details should be abstracted from them, and what new mechanisms might be used to make explanations understandable to them, we envision a future of human-centered XAI that combines participatory design methods into XAI development. For this to be accomplished, researchers need to: (1) spend substantial amounts of time in the field to fully understand the context as well as the existing strengths and resources within the community the AI tools will be deployed in, (2) use participatory methods that treat community members as active agents in the design process, and (3) engage in evaluations of XAI methods *in situ* with end-users instead of relying on crowdsourcing from general populations present on platforms such as Amazon Mechanical Turk. In line with work by Eiband et al. [51] that developed a participatory design method to shift users' mental models when interacting with AI-enabled systems, by understanding the gaps between users' mental model of a system and how developers intend for the respective system to work, users can inform what aspects need to be explained and how they can be explained. It is equally important to study the implications of these methods both qualitatively and empirically. Continuing to conduct empirical studies with a variety of stakeholders is essential in providing researchers with the ability to make informed design approaches toward explainable AI that caters to a wide range of diverse end-users. Recently developed tools like the OpenXAI library, [3] which provides an open-source framework to evaluate and benchmark post-hoc explanation methods, and HIVE, [84] which targets human users to evaluate the utility of computer vision XAI methods, show promise in advancing such empirical work.

Research continually emphasizes the need for XAI to center stakeholders while acknowledging their distinct needs [17, 22, 59, 94, 124, 135, 158], however current approaches to XAI methods drastically fail to do so for populations and contexts in the Global South. For example, as most AI/ML vocabulary is presented in English, considerations regarding the choice of language also have to be explored. In the case of our CHW participants, the language barriers that existed when we translated the textual aspects of LIME and SHAP (colorbar labels, image labels) from English to Hindi may also hold the same for other languages not commonly represented in ML. For example, संभावना ("sambhaavna") the closest word to "predicted" in Hindi, also translates literally to

"possibility." One CHW (P10) in our study interpreted this to mean that the app probe provides an uncertain diagnosis of jaundice and the *actual* diagnosis would be given by a doctor. Researchers could potentially fill this void by collaborating across regions to create a working dictionary of X/AI jargons with simple definitions and translations into local languages and dialects. Prior work has created a dictionary of AI terms but such dictionaries date to the 1980s [142], 1990s [166, 175], and only focus on translations to Western languages like English, German, French, and Italian [166]. Such an X/AI dictionary would expand upon the efforts by Skočaj et al.[151] that built a dictionary of AI terms in Slovenian and a series of YouTube videos by Wuraola Oyewusi [129] introducing the concepts of AI in Yoruba, a language widely spoken in Nigeria. The establishment of such resources would enable future developers of XAI toolkits and libraries to incorporate these translated terms, potentially serving a wider range of linguistically diverse users in non-Western contexts.

5.3 Scaffolding explanations

The major frustrations relayed by our participants primarily focused on their lack of understanding regarding basic AI tenets and mathematical concepts. As AI-driven tools are integrated into everyday workflows of users like CHWs, who not only lack AI know-how but also have limited domain expertise, new explainability methods that simplify complex math and statistical terms is only the first step. Even more important is to understand how to explain AI to those who not only have low AI knowledge, but might also be tasked to explain the inner workings of AI to others, as seen in our findings and in work by Tonekaboni et al. [160] where clinicians view explanations as a potential method to justify medical decision-making guided by AI systems.

Given low levels of AI knowledge and exposure among CHWs, it cannot be expected that CHWs will develop capabilities to understand intricate explanations themselves. It is critical to provide scaffolding structures that support them in understanding how AI operates under the hood and what realistic expectations they should have when working with AI-driven tools. Many of our participants expressed that getting training on understanding explanations would improve their utility for real-world use. For example, when users are exposed to apps or computer programs for the first time, they receive explicit training that is offered in the form of short-form tutorials or guided interactions. We believe that similar training should be incorporated to guide users when interacting with AI tools and XAI interfaces.

There is a need to work with CHWs to understand the type of support and training they may need to understand explanations. Compared with traditional approaches taken in developing and deploying ML systems, the HCI community strongly advocates for using communities' existing methods and frameworks and building on them to expand to new areas rather than using one-size-fits-all approaches [1, 55, 67, 103, 149]. For example, our participants found it particularly hard to understand the concept of uncertainty as it directly conflicted with their existing mental models of the devices they use daily in their work. The concept of AI being a *prediction* rather than a *measurement* such as a weight or a temperature skewed their understanding of explanations. To bridge this fundamental difference in thinking, we can develop support programs that build upon their existing knowledge in ways that make sense for the community, for instance, by using contextually-relevant analogies and examples or by turning to their prior experiences with medical uncertainty where equipment failure and complex patient symptoms can lead to indecision when treating patients [64, 83, 173].

5.4 Explaining AI through Visualizations

Within our work on incorporating visualization-based XAI methods, we found that the participant mental models of colors often impacted their interpretation of how the app detected jaundice.

Other graphical and textual features like boxes, reference images, labels, and legends were also widely misinterpreted. Further, we found that our efforts to remove color improved understanding and allowed CHWs to rely more on their domain knowledge to interpret the explanations. Such slight changes show promise in aiding the understanding of XAI for diseases similar to jaundice which are impacted by discrete features such as color. However, there is more work required to make the task of explaining predictions from AI systems to novice technology users much easier.

There is a large body of research within HCI that discusses the use of text-free interfaces for novice and low-literate technology users [29, 105, 133] and ways to effectively use color in user interfaces [87, 101, 102, 106]. Despite this, there is a lack of guidelines for stipulating the best use of color and visualizations in graphical approaches to XAI, leaving plenty of room for researchers to incorporate HCI principles into this domain. There is also a burgeoning amount of literature focused on examining the impact of visualizations in XAI [34, 159, 179]. However, this research shows mixed results for the ability of visual-based XAI approaches to improve user trust and understanding of model predictions. With this in mind, we call for more research into radical design approaches of visualizations within XAI. The emergence of work focusing on developing novel visualization-based XAI methods [19, 61, 68, 127, 165] shows that there is potential for new XAI modalities that combine informative visualizations and layperson-interpretable explanations in verbal or text form. A solid example of this is an approach by Biran and McKeown [19] that uses natural language generation to produce simple explanations combined with graphic visualizations to explain the prediction and the feature contributions. To improve users' understanding, features are color-coded with their accompanying explanation. Their evaluation demonstrates that this method helps users understand the accuracy of a prediction and increases their satisfaction with predictions, showing promise for more work within this space.

One of the challenges with visualization-based explainability methods is the higher complexity of information shown. While removing color from our explanations drastically improved our participants' understanding of the predictions, we cannot conjecture that colors shouldn't be used in explanations tailored to novice technology users. Additionally, given the focus of this work on just two explanation methods, it is also hard to determine if the failure is within the way *we* chose to visualize the explanations, if fundamentally the information associated with jaundice diagnosis being explained is too complicated, or if LIME and SHAP themselves are too complicated. However, there are not straightforward ways to convey such nuanced information in a simple manner even through text-based explanations. We call for more research into how other XAI strategies, other than LIME or SHAP, can be implemented into end-user-friendly visual XAI. By leveraging these advances in tandem, along with incorporating the design recommendations discussed previously, we can unlock the potential for a new era of novel XAI methods that meet the needs of diverse users globally.

6 CONCLUSION

This paper details a qualitative study examining how CHWs, who are increasingly expected to cooperatively work with AI-driven tools, engage with and perceive AI explainability methods. Additionally, our work explores how such methods can be redesigned to improve CHWs' AI understanding. Using an AI-based jaundice diagnostic app as motivation, we uncover conflicts between CHWs' X/AI understanding, mental models of disease diagnosis, color associations, and perceived benefits of having access to explanations. We conclude by discussing (1) what elements of AI need to be explainable for end users, (2) design considerations for improving access to XAI in the Global South, (3) potential ways to support non-technical users in interpreting XAI, and (4) how visualizations can be leveraged to improve current XAI methods. Our findings highlight opportunities for new domains of XAI research to account for socially, culturally, and linguistically diverse users

with low levels of AI knowledge and augment their collaboration with AI-driven tools. The existence of community health work in regions beyond India and the use of XAI in domains beyond medicine indicates a possibility for our findings to generalize beyond CHWs in rural India. With this in mind, subsequent studies will be required to explore how this work can be relevant in other domains.

REFERENCES

[1] Jose Abdelnour-Nocera, Torkil Clemmensen, and Masaaki Kurosu. 2013. Reframing HCI through local and indigenous perspectives. *International Journal of Human-Computer Interaction* 29, 4 (2013), 201–204.

[2] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. *Advances in neural information processing systems* 31 (2018).

[3] Chirag Agarwal, Eshika Saxena, Satyapriya Krishna, Martin Pawelczyk, Nari Johnson, Isha Puri, Marinka Zitnik, and Himabindu Lakkaraju. 2022. OpenXAI: Towards a Transparent Evaluation of Model Explanations. *arXiv preprint arXiv:2206.11104* (2022).

[4] Wadhwani AI. 2020. Cough Against Covid. Retrieved November 28, 2022 from https://www.wadhwaniai.org/programs/cough-against-covid/

[5] Wadhwani AI. 2020. Newborn Anthropometry. Retrieved November 28, 2022 from https://www.wadhwaniai.org/programs/newborn-anthropometry/

[6] Maximilian Alber, Sebastian Lapuschkin, Philipp Seegerer, Miriam Hägele, Kristof T Schütt, Grégoire Montavon, Wojciech Samek, Klaus-Robert Müller, Sven Dähne, and Pieter-Jan Kindermans. 2019. iNNvestigate neural networks! *J. Mach. Learn. Res.* 20, 93 (2019), 1–8.

[7] Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze. 2020. Evaluating saliency map explanations for convolutional neural networks: a user study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 275–285.

[8] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.

[9] Ariful Islam Anik and Andrea Bunt. 2021. Data-centric explanations: explaining training data of machine learning systems to promote transparency. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[10] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115.

[11] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[12] Abdullah H Baqui, Shams El-Arifeen, Gary L Darmstadt, Saifuddin Ahmed, Emma K Williams, Habibur R Seraji, Ishtiaq Mannan, Syed M Rahman, Rasheduzzaman Shah, Samir K Saha, et al. 2008. Effect of community-based newborn-care intervention package implemented through two service-delivery strategies in Sylhet district, Bangladesh: a cluster-randomised controlled trial. *The lancet* 371, 9628 (2008), 1936–1944.

[13] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M Martinez, and Seth D Pollak. 2019. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological science in the public interest* 20, 1 (2019), 1–68.

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

[15] Adrien Bennetot, Jean-Luc Laurent, Raja Chatila, and Natalia Díaz-Rodríguez. 2019. Towards explainable neural-symbolic visual reasoning. *arXiv preprint arXiv:1909.09065* (2019).

[16] Umang Bhatt, McKane Andrus, Adrian Weller, and Alice Xiang. 2020. Machine learning explainability for external stakeholders. *arXiv preprint arXiv:2007.05408* (2020).

[17] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José MF Moura, and Peter Eckersley. 2020. Explainable machine learning in deployment. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 648–657.

[18] Przemysław Biecek. 2018. DALEX: explainers for complex predictive models in R. *The Journal of Machine Learning Research* 19, 1 (2018), 3245–3249.

[19] Or Biran and Kathleen R McKeown. 2017. Human-Centric Justification of Machine Learning Predictions.. In *IJCAI*, Vol. 2017. 1461–1467.

[20] Claus Bossen and Martin Foss. 2016. The collaborative work of hospital porters: Accountability, visibility and configurations of work. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 965–979.

[21] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[22] Andrea Brennen. 2020. What Do People Really Want When They Say They Want" Explainable AI?" We Asked 60 Stakeholders.. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–7.

[23] Emma Brunskill and Neal Lesh. 2010. Routing for rural health: optimizing community health worker visit schedules. In *2010 AAAI Spring Symposium Series*.

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

[25] Carrie J Cai, Jonas Jongejan, and Jess Holbrook. 2019. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th international conference on intelligent user interfaces*. 258–262.

[26] David Canfield Smith, Charles Irby, Ralph Kimball, and Bill Verplank. 1982. Designing the Star User Interface. Retrieved November 8, 2022 from https://www.tech-insider.org/star/research/acrobat/8204.pdf

[27] Elizabeth Charters. 2003. The Use of Think-aloud Methods in Qualitative Research An Introduction to Think-aloud Methods. *Brock Education Journal* 12, 2 (July 2003). https://doi.org/10.26522/brocked.v12i2.38 Number: 2.

[28] Jianbo Chen, Le Song, Martin J Wainwright, and Michael I Jordan. 2018. L-Shapley and C-Shapley: Efficient Model Interpretation for Structured Data. In *International Conference on Learning Representations*.

[29] Jung-Wei Chen and Jiajie Zhang. 2007. Comparing text-based and graphic user interfaces for novice and expert users. In *AMIA annual symposium proceedings*, Vol. 2007. American Medical Informatics Association, 125.

[30] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. 2019. Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.

[31] Minseok Cho, Gyeongbok Lee, and Seung-won Hwang. 2019. Explanatory and actionable debugging for machine learning: A tableqa demonstration. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1333–1336.

[32] Michael Chromik and Andreas Butz. 2021. Human-XAI interaction: a review and design principles for explanation user interfaces. In *IFIP Conference on Human-Computer Interaction*. Springer, 619–640.

[33] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. 2021. I think i get your point, AI! the illusion of explanatory depth in explainable AI. In *26th International Conference on Intelligent User Interfaces*. 307–317.

[34] Eric Chu, Deb Roy, and Jacob Andreas. 2020. Are visual explanations useful? a case study in model-in-the-loop prediction. *arXiv preprint arXiv:2007.12248* (2020).

[35] Chia-Fang Chung, Kristin Dew, Allison Cole, Jasmine Zia, James Fogarty, Julie A Kientz, and Sean A Munson. 2016. Boundary negotiating artifacts in personal informatics: patient-provider collaboration with patient-generated data. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*. 770–786.

[36] Danielle Keats Citron and Frank Pasquale. 2014. The scored society: Due process for automated predictions. *Wash. L. Rev.* 89 (2014), 1.

[37] Lilian De Greef, Mayank Goel, Min Joon Seo, Eric C Larson, James W Stout, James A Taylor, and Shwetak N Patel. 2014. Bilicam: using mobile phones to monitor newborn jaundice. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 331–342.

[38] Nicola Dell, Jessica Crawford, Nathan Breit, Timóteo Chaluco, Aida Coelho, Joseph McCord, and Gaetano Borriello. 2013. Integrating ODK Scan into the Community Health Worker Supply Chain in Mozambique. In *Proceedings of the Sixth International Conference on Information and Communication Technologies and Development: Full Papers - Volume 1* (Cape Town, South Africa) *(ICTD '13)*. Association for Computing Machinery, New York, NY, USA, 228–237. https://doi.org/10.1145/2516604.2516611

[39] Nicola Dell and Neha Kumar. 2016. The ins and outs of HCI for development. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 2220–2232.

[40] Brian DeRenzi, Nicola Dell, Jeremy Wacksman, Scott Lee, and Neal Lesh. 2017. Supporting community health workers in India through voice-and web-based feedback. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. 2770–2781.

[41] Brian DeRenzi, Neal Lesh, Tapan Parikh, Clayton Sims, Werner Maokla, Mwajuma Chemba, Yuna Hamisi, David S hellenberg, Marc Mitchell, and Gaetano Borriello. 2008. E-IMCI: Improving pediatric health care in low-income

countries. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 753–762.
[42] Brian DeRenzi, Jeremy Wacksman, Nicola Dell, Scott Lee, Neal Lesh, Gaetano Borriello, and Andrew Ellner. 2016. Closing the feedback Loop: A 12-month evaluation of ASTA, a self-tracking application for ASHAs. In *Proceedings of the Eighth International Conference on Information and Communication Technologies and Development*. 1–10.
[43] Skater developers and contributors. 2017. Skater. Retrieved October 19, 2022 from https://github.com/oracle/Skater
[44] Shipi Dhanorkar, Christine T Wolf, Kun Qian, Anbang Xu, Lucian Popa, and Yunyao Li. 2021. Who needs to know what, when?: Broadening the Explainable AI (XAI) Design Space by Looking at Explanations Across the AI Lifecycle. In *Designing Interactive Systems Conference 2021*. 1591–1602.
[45] Aparna Dhinakaran. 2022. A Look Into Global, Cohort and Local Model Explainability. Retrieved October 31, 2022 from https://towardsdatascience.com/a-look-into-global-cohort-and-local-model-explainability-973bd449969f
[46] Jonathan Dodge, Q Vera Liao, Yunfeng Zhang, Rachel KE Bellamy, and Casey Dugan. 2019. Explaining models: an empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th international conference on intelligent user interfaces*. 275–285.
[47] Jeff Druce, Michael Harradon, and James Tittle. 2021. Explainable Artificial Intelligence (XAI) for Increasing User Trust in Deep Reinforcement Learning Driven Autonomous Systems. *arXiv preprint arXiv:2106.03775* (2021).
[48] Upol Ehsan, Q Vera Liao, Michael Muller, Mark O Riedl, and Justin D Weisz. 2021. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–19.
[49] Upol Ehsan and Mark O Riedl. 2020. Human-centered explainable ai: Towards a reflective sociotechnical approach. In *International Conference on Human-Computer Interaction*. Springer, 449–466.
[50] Upol Ehsan and Mark O Riedl. 2021. Explainability pitfalls: Beyond dark patterns in explainable AI. *arXiv preprint arXiv:2109.12480* (2021).
[51] Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. 2018. Bringing transparency design into practice. In *23rd international conference on intelligent user interfaces*. 211–223.
[52] Márcia Cristina Rodrigues Fausto, Ligia Giovanella, Maria Helena Magalhães de Mendonça, Patty Fidelis de Almeida, Sarah Escorel, Carla Lourenço Tavares de Andrade, and Maria Inês Carsalade Martins. 2011. The work of community health workers in major cities in Brazil: mediation, community action, and health care. *The Journal of ambulatory care management* 34, 4 (2011), 339–353.
[53] The Institute for Ethical AI & ML. 2021. XAI - An eXplainability toolbox for machine learning. Retrieved October 21, 2022 from https://github.com/EthicalML/xai
[54] Maximilian Förster, Mathias Klier, Kilian Kluge, and Irina Sigler. 2020. Fostering human agency: a process for the design of user-centric XAI systems. (2020).
[55] Isabela Gasparini, Marcelo S Pimenta, and José Palazzo M De Oliveira. 2011. Vive la différence! a survey of cultural-aware issues in HCI. In *Proceedings of the 10th Brazilian Symposium on Human Factors in Computing Systems and the 5th Latin American Conference on Human-Computer Interaction*. 13–22.
[56] Bill Gaver, Tony Dunne, and Elena Pacenti. 1999. Design: cultural probes. *interactions* 6, 1 (1999), 21–29.
[57] Rob Geada, Tommaso Teofili, Rui Vieira, Rebecca Whitworth, and Daniele Zonca. 2021. TrustyAI Explainability Toolkit. *arXiv preprint arXiv:2104.12717* (2021).
[58] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.
[59] Julie Gerlings, Arisa Shollo, and Ioanna Constantiou. 2020. Reviewing the need for explainable artificial intelligence (xAI). *arXiv preprint arXiv:2012.01007* (2020).
[60] Google. 2020. AI Experiments - Experiments with Google. Retrieved October 19, 2022 from https://experiments.withgoogle.com/collection/ai
[61] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Counterfactual visual explanations. In *International Conference on Machine Learning*. PMLR, 2376–2384.
[62] Hani Hagras. 2018. Toward human-understandable, explainable AI. *Computer* 51, 9 (2018), 28–36.
[63] Andy Haines, David Sanders, Uta Lehmann, Alexander K Rowe, Joy E Lawn, Steve Jan, Damian G Walker, and Zulfiqar Bhutta. 2007. Achieving child survival goals: potential contribution of community health workers. *The lancet* 369, 9579 (2007), 2121–2131.
[64] Paul KJ Han, William MP Klein, and Neeraj K Arora. 2011. Varieties of uncertainty in health care: a conceptual taxonomy. *Medical Decision Making* 31, 6 (2011), 828–838.
[65] Carl Hartung, Adam Lerer, Yaw Anokwa, Clint Tseng, Waylon Brunette, and Gaetano Borriello. 2010. Open data kit: tools to build information services for developing regions. In *Proceedings of the 4th ACM/IEEE international conference on information and communication technologies and development*. 1–12.
[66] Anna Hedström, Leander Weber, Dilyara Bareeva, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina M-C Höhne. 2022. Quantus: an explainable AI toolkit for responsible evaluation of neural network explanations.

arXiv preprint arXiv:2202.06861 (2022).

[67] Rüdiger Heimgärtner. 2013. Intercultural User Interface Design–Culture-Centered HCI Design–Cross-Cultural User Interface Design: Different Terminology or Different Approaches?. In *International Conference of Design, User Experience, and Usability*. Springer, 62–71.

[68] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. 2016. Generating visual explanations. In *European conference on computer vision*. Springer, 3–19.

[69] Michael Hind. 2019. Explaining explainable AI. *XRDS: Crossroads, The ACM Magazine for Students* 25, 3 (2019), 16–19.

[70] Robert R Hoffman, Gary Klein, and Shane T Mueller. 2018. Explaining explanation for "explainable AI". In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 62. SAGE Publications Sage CA: Los Angeles, CA, 197–201.

[71] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).

[72] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M Drucker. 2019. Gamut: A design probe to understand how data scientists understand machine learning models. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.

[73] Steven Houben, Mads Frost, and Jakob E Bardram. 2015. Collaborative affordances of hybrid patient record technologies in medical work. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. 785–797.

[74] Brian Hu, Paul Tunison, Bhavan Vasu, Nitesh Menon, Roddy Collins, and Anthony Hoogs. 2021. XAITK: The explainable AI toolkit. *Applied AI Letters* 2, 4 (2021), e40.

[75] Sami Hulkko, Tuuli Mattelmäki, Katja Virtanen, and Turkka Keinonen. 2004. Mobile probes. In *Proceedings of the third Nordic conference on Human-computer interaction*. 43–51.

[76] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B. Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, Nicolas Roussel, and Björn Eiderbäck. 2003. Technology probes: inspiring design for and with families. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '03)*. Association for Computing Machinery, New York, NY, USA, 17–24. https://doi.org/10.1145/642611.642616

[77] Azra Ismail and Neha Kumar. 2018. Engaging solidarity in data collection practices for community health. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–24.

[78] Azra Ismail and Neha Kumar. 2021. AI in global health: the view from the front lines. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–21.

[79] WEINA Jin, JIANYU Fan, D Gromala, P Pasquier, and G Hamarneh. 2021. EUCA: the End-User-Centered Explainable AI Framework. *arXiv preprint arXiv:2102.02437* (2021).

[80] Hyungsik Jung and Youngrock Oh. 2021. Towards Better Explanations of Class Activation Mapping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1336–1344.

[81] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–14.

[82] Kimia Kiani, George Cui, Andrea Bunt, Joanna McGrenere, and Parmit K. Chilana. 2019. Beyond "One-Size-Fits-All": Understanding the Diversity in How Software Newcomers Discover and Make Use of Help Resources. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300570

[83] Kangmoon Kim and Young-Mee Lee. 2018. Understanding uncertainty in medicine: concepts and implications in medical education. *Korean journal of medical education* 30, 3 (2018), 181.

[84] Sunnie SY Kim, Nicole Meister, Vikram V Ramaswamy, Ruth Fong, and Olga Russakovsky. 2021. Hive: evaluating the human interpretability of visual explanations. *arXiv preprint arXiv:2112.03184* (2021).

[85] Janis Klaise, Arnaud Van Looveren, Giovanni Vacanti, and Alexandru Coca. 2021. Alibi Explain: Algorithms for Explaining Machine Learning Models. *J. Mach. Learn. Res.* 22 (2021), 181–1.

[86] Beth E Kolko, Alexis Hope, Waylon Brunette, Karen Saville, Wayne Gerard, Michael Kawooya, and Robert Nathan. 2012. Adapting collaborative radiological practice to low-resource environments. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. 97–106.

[87] Irina Kondratova and Ilia Goldfarb. 2006. Cultural interface design: Global colors study. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer, 926–934.

[88] Satyapriya Krishna, Tessa Han, Alex Gu, Javin Pombra, Shahin Jabbari, Steven Wu, and Himabindu Lakkaraju. 2022. The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective. *arXiv preprint arXiv:2202.01602* (2022).

[89] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*. 126–137.

[90] Neha Kumar and Richard J Anderson. 2015. Mobile phones for maternal health in rural India. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 427–436.

[91] Vishwajeet Kumar, Saroj Mohanty, Aarti Kumar, Rajendra P Misra, Mathuram Santosham, Shally Awasthi, Abdullah H Baqui, Pramod Singh, Vivek Singh, Ramesh C Ahuja, et al. 2008. Effect of community-based behaviour change management on neonatal mortality in Shivgarh, Uttar Pradesh, India: a cluster-randomised controlled trial. *The Lancet* 372, 9644 (2008), 1151–1162.

[92] Diana S Kusunoki, Aleksandra Sarcevic, Zhan Zhang, and Randall S Burd. 2013. Understanding visual attention of teams in dynamic medical settings through vital signs monitor use. In *Proceedings of the 2013 conference on Computer supported cooperative work*. 527–540.

[93] Himabindu Lakkaraju and Osbert Bastani. 2020. " How do I fool you?" Manipulating User Trust via Misleading Black Box Explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 79–85.

[94] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. 2021. What do we want from Explainable Artificial Intelligence (XAI)?–A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence* 296 (2021), 103473.

[95] Kwang Hee Lee, Chaewon Park, Junghyun Oh, and Nojun Kwak. 2021. LFI-CAM: Learning Feature Importance for Better Visual Explanation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1355–1363.

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

[97] Sebastian Linxen, Christian Sturm, Florian Brühlmann, Vincent Cassau, Klaus Opwis, and Katharina Reinecke. 2021. How WEIRD is CHI?. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 143, 14 pages. https://doi.org/10.1145/3411764.3445488

[98] Pei-Ju Liu, James M Laffey, and Karen R Cox. 2008. Operationalization of technology use and cooperation in CSCW. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*. 505–514.

[99] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*. 4768–4777.

[100] B Mane Abhay and V Khandekar Sanjay. 2014. Strengthening primary health care through Asha Workers: a novel approach in India. *Primary Health Care* 4, 149 (2014), 2167–1079.

[101] Aaron Marcus. 1995. Principles of effective visual communication for graphical user interface design. In *Readings in human–computer interaction*. Elsevier, 425–441.

[102] Aaron Marcus, William B Cowan, and Wanda Smith. 1989. Color in user interface design: functionally and aesthetics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 25–27.

[103] Aaron Marcus and Emilie W Gould. 2012. Globalization, localization, and cross-cultural user-interface design. , 341–366 pages.

[104] Bernardo Mariano. 2020. Towards a global strategy on digital health. *Bulletin of the World Health Organization* 98, 4 (2020), 231.

[105] Indrani Medhi, Aman Sagar, and Kentaro Toyama. 2006. Text-free user interfaces for illiterate and semi-literate users. In *2006 international conference on information and communication technologies and development*. IEEE, 72–82.

[106] Barbara J Meier. 1988. ACE: a color expert system for user interface design. In *Proceedings of the 1st annual ACM SIGGRAPH symposium on User Interface Software*. 117–128.

[107] Tim Miller. 2019. " But why?" Understanding explainable artificial intelligence. *XRDS: Crossroads, The ACM Magazine for Students* 25, 3 (2019), 20–25.

[108] Joy Ming, Srujana Kamath, Elizabeth Kuo, Madeline Sterling, Nicola Dell, and Aditya Vashistha. 2022. Invisible Work in Two Frontline Health Contexts. In *ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies (COMPASS)* (Seattle, WA, USA) *(COMPASS '22)*. Association for Computing Machinery, New York, NY, USA, 139–151. https://doi.org/10.1145/3530190.3534814

[109] 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*. 220–229.

[110] Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining explanations in AI. In *Proceedings of the conference on fairness, accountability, and transparency*. 279–288.

[111] Sina Mohseni, Niloofar Zarei, and Eric D Ragan. 2021. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 11, 3-4 (2021), 1–45.

[112] Maletsabisa Molapo, Melissa Densmore, and Brian DeRenzi. 2017. Video consumption patterns for first time smartphone users: Community health workers in Lesotho. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 6159–6170.

[113] Maletsabisa Molapo, Melissa Densmore, and Limpho Morie. 2016. Designing with Community Health Workers: Enabling Productive Participation Through Exploration. In *Proceedings of the First African Conference on Human Computer Interaction (AfriCHI'16)*. Association for Computing Machinery, New York, NY, USA, 58–68. https://doi.org/10.1145/2998581.2998589

[114] Christoph Molnar. 2020. *Interpretable machine learning*.

[115] Satya M Muddamsetty, NS Jahromi Mohammad, and Thomas B Moeslund. 2020. Sidu: Similarity difference and uniqueness method for explainable ai. In *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 3269–3273.

[116] Shane T Mueller, Elizabeth S Veinott, Robert R Hoffman, Gary Klein, Lamia Alam, Tauseef Mamun, and William J Clancey. 2021. Principles of explanation in human-AI systems. *arXiv preprint arXiv:2102.04972* (2021).

[117] T Nathan Mundhenk, Barry Y Chen, and Gerald Friedland. 2019. Efficient saliency maps for Explainable AI. *arXiv preprint arXiv:1911.11293* (2019).

[118] Grace W Mwai, Gitau Mburu, Kwasi Torpey, Peter Frost, Nathan Ford, and Janet Seeley. 2013. Role and outcomes of community health workers in HIV care in sub-Saharan Africa: a systematic review. *Journal of the International AIDS Society* 16, 1 (2013), 18586.

[119] Siddharth Nishtala, Harshavardhan Kamarthi, Divy Thakkar, Dhyanesh Narayanan, Anirudh Grama, Aparna Hegde, Ramesh Padmanabhan, Neha Madhiwalla, Suresh Chaudhary, Balaraman Ravindran, et al. 2020. Missed calls, Automated Calls and Health Support: Using AI to improve maternal health outcomes by increasing program engagement. *arXiv preprint arXiv:2006.07590* (2020).

[120] Greg Noone. 2022. Emotion recognition is mostly ineffective. Why are companies still investing in it? Retrieved October 1, 2022 from https://techmonitor.ai/technology/emerging-technology/emotion-recognition

[121] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. 2019. Interpretml: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223* (2019).

[122] Fabian Okeke, Lucas Nene, Anne Muthee, Stephen Odindo, Dianna Kane, Isaac Holeman, and Nicola Dell. 2019. Opportunities and challenges in connecting care recipients to the community health feedback loop. In *Proceedings of the Tenth International Conference on Information and Communication Technologies and Development*. 1–11.

[123] Chinasa T Okolo. 2022. Optimizing human-centered AI for healthcare in the Global South. *Patterns* (2022), 100421.

[124] Chinasa T. Okolo, Nicola Dell, and Aditya Vashistha. 2022. Making AI Explainable in the Global South: A Systematic Review. In *ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies (COMPASS) (COMPASS '22)*. Association for Computing Machinery, 439–452.

[125] Chinasa T Okolo, Srujana Kamath, Nicola Dell, and Aditya Vashistha. 2021. "It cannot do all of my work": community health worker perceptions of AI-enabled mobile health applications in rural India. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–20.

[126] Abimbola Olaniran, Barbara Madaj, Sarah Bar-Zev, and Nynke van den Broek. 2019. The roles of community health workers who provide maternal and newborn health services: case studies from Africa and Asia. *BMJ global health* 4, 4 (2019), e001388.

[127] Ahmet Haydar Örnek and Murat Ceylan. 2022. A Novel Approach for Visualization of Class Activation Maps with Reduced Dimensions. In *2022 Innovations in Intelligent Systems and Applications Conference (ASYU)*. IEEE, 1–5.

[128] Ayomide Owoyemi, Joshua Owoyemi, Adenekan Osiyemi, and Andy Boyd. 2020. Artificial intelligence for healthcare in Africa. *Frontiers in Digital Health* 2 (2020), 6.

[129] Wuraola Oyewusi. 2022. Learn Tech Concepts in Yoruba Language. Retrieved November 15, 2022 from https://www.youtube.com/playlist?list=PLwYqTHZgNtLwPw2PtydKvb1fCyRqodBaf

[130] James O'Donovan, Ken Kahn, MacKenzie MacRae, Allan Saul Namanda, Rebecca Hamala, Ken Kabali, Anne Geniets, Alice Lakati, Simon M Mbae, and Niall Winters. 2022. Analysing 3429 digital supervisory interactions between Community Health Workers in Uganda and Kenya: the development, testing and validation of an open access predictive machine learning web app. *Human resources for health* 20, 1 (2022), 1–8.

[131] Joyojeet Pal, Anjuli Dasika, Ahmad Hasan, Jackie Wolf, Nick Reid, Vaishnav Kameswaran, Purva Yardi, Allyson Mackay, Abram Wagner, Bhramar Mukherjee, et al. 2017. Changing data practices for community health workers: Introducing digital data collection in West Bengal, India. In *Proceedings of the Ninth International Conference on Information and Communication Technologies and Development*. 1–12.

[132] Chunjong Park, Alex Mariakakis, Jane Yang, Diego Lassala, Yasamba Djiguiba, Youssouf Keita, Hawa Diarra, Beatrice Wasunna, Fatou Fall, Marème Soda Gaye, et al. 2020. Supporting Smartphone-Based Image Capture of Rapid Diagnostic Tests in Low-Resource Settings. In *Proceedings of the 2020 International Conference on Information and Communication Technologies and Development*. 1–11.

[133] Madeline Plauché and Udhyakumar Nallasamy. 2007. Speech interfaces for equitable access to information technology. *Information Technologies & International Development* 4, 1 (2007), pp–69.

[134] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–52.

[135] Alun Preece, Dan Harborne, Dave Braines, Richard Tomsett, and Supriyo Chakraborty. 2018. Stakeholders in explainable AI. *arXiv preprint arXiv:1810.00184* (2018).

[136] Divya Ramachandran, John Canny, Prabhu Dutta Das, and Edward Cutrell. 2010. Mobile-izing health workers in rural India. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1889–1898.

[137] General Data Protection Regulation. 2018. General data protection regulation (GDPR). *Intersoft Consulting, Accessed in October* 24, 1 (2018).

[138] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[139] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[140] Mireia Ribera and Agata Lapedriza. 2019. Can we do better explanations? A proposal of user-centered explainable AI.. In *IUI Workshops*, Vol. 2327. 38.

[141] Sarah M Rodrigues, Anil Kanduri, Adeline Nyamathi, Nikil Dutt, Pramod Khargonekar, and Amir M Rahmani. 2022. Digital Health–Enabled Community-Centered Care: Scalable Model to Empower Future Community Health Workers Using Human-in-the-Loop Artificial Intelligence. *JMIR formative research* 6, 4 (2022), e29535.

[142] Jerry Martin Rosenberg. 1986. Dictionary of artificial intelligence and robotics. (1986).

[143] Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. 2021. Re-Imagining Algorithmic Fairness in India and Beyond. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 315–328. https://doi.org/10.1145/3442188.3445896

[144] Helen Schneider, Dickson Okello, and Uta Lehmann. 2016. The global pendulum swing towards community health workers in low-and middle-income countries: a scoping review of trends, geographical distribution and programmatic orientations, 2005 to 2014. *Human resources for health* 14, 1 (2016), 1–12.

[145] Johanes Schneider and Joshua Handali. 2019. Personalized explanation in machine learning: A conceptualization. *arXiv preprint arXiv:1901.00770* (2019).

[146] Nina Schwalbe and Brian Wahl. 2020. Artificial intelligence and the future of global health. *The Lancet* 395, 10236 (2020), 1579–1586.

[147] Kerry Scott, SW Beckham, Margaret Gross, George Pariyo, Krishna D Rao, Giorgio Cometto, and Henry B Perry. 2018. What do we know about community-based health worker programs? A systematic review of existing reviews on community health workers. *Human resources for health* 16, 1 (2018), 1–17.

[148] Hong Shen, Haojian Jin, Ángel Alexander Cabrera, Adam Perer, Haiyi Zhu, and Jason I Hong. 2020. Designing Alternative Representations of Confusion Matrices to Support Non-Expert Public Understanding of Algorithm Performance. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–22.

[149] Siu-Tsen Shen, Martin Woolley, and Stephen Prior. 2006. Towards culture-centred design. *Interacting with computers* 18, 4 (2006), 820–852.

[150] Line Silsand and Gunnar Ellingsen. 2016. Complex decision-making in clinical practice. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 993–1004.

[151] Danijel Skočaj, Damjan Strnad, Marko Robnik Šikonja, Vladimir Batagelj, Ivan Bratko, Matjaž Divjak, Peter Rogelj, Marko Bizjak, Boštjan Slivnik, and Tanja Fajfar. 2022. Terminological dictionary of artificial intelligence. (2022).

[152] Daniel Smilkov and Shan Carter. 2020. Tensorflow-Neural Network Playground. Retrieved October 19, 2022 from https://playground.tensorflow.org/

[153] Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Daniel S Weld, and Leah Findlater. 2020. No explainability without accountability: An empirical study of explanations and feedback in interactive ml. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[154] Aaron Springer and Steve Whittaker. 2019. Progressive disclosure: empirically motivated approaches to designing effective transparency. In *Proceedings of the 24th international conference on intelligent user interfaces*. 107–120.

[155] S. Sue. 1999. Science, ethnicity, and bias: where have we gone wrong? *The American Psychologist* 54, 12 (Dec. 1999), 1070–1077. https://doi.org/10.1037/0003-066x.54.12.1070

[156] Yunjia Sun. 2016. Novice-Centric Visualizations for Machine Learning. (2016).

[157] Yunjia Sun, Edward Lank, and Michael Terry. 2017. Label-and-Learn: Visualizing the Likelihood of Machine Learning Classifier's Success During Data Labeling. (2017), 523–534.

[158] Harini Suresh, Steven R Gomez, Kevin K Nam, and Arvind Satyanarayan. 2021. Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[159] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. 2021. Visual, textual or hybrid: the effect of user expertise on different explanations. In *26th International Conference on Intelligent User Interfaces*. 109–119.

[160] Sana Tonekaboni, Shalmali Joshi, Melissa D McCradden, and Anna Goldenberg. 2019. What clinicians want: contextualizing explainable machine learning for clinical end use. In *Machine learning for healthcare conference*. PMLR, 359–380.

[161] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael Bernstein, and Ranjay Krishna. 2022. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *arXiv preprint arXiv:2212.06823* (2022).

[162] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Ranjay Krishna, Tobias Gerstenberg, and Michael S. Bernstein. 2022. When Do XAI Methods Work? A Cost-Benefit Approach to Human-AI Collaboration. (2022).

[163] Aditya Vashistha, Neha Kumar, Anil Mishra, and Richard Anderson. 2016. Mobile video dissemination for community health. In *Proceedings of the Eighth International Conference on Information and Communication Technologies and Development*. 1–11.

[164] Aditya Vashistha, Neha Kumar, Anil Mishra, and Richard Anderson. 2017. Examining localization approaches for community health. In *Proceedings of the 2017 Conference on Designing Interactive Systems*. 357–368.

[165] Tom Vermeire, Dieter Brughmans, Sofie Goethals, Raphael Mazzine Barbossa de Oliveira, and David Martens. 2022. Explainable image classification with evidence counterfactual. *Pattern Analysis and Applications* (2022), 1–21.

[166] Otto Vollnhals. 1992. *A multilingual dictionary of artificial intelligence: English, German, French, Spanish, Italian*. Psychology Press.

[167] Brian Wahl, Aline Cossy-Gantner, Stefan Germann, and Nina R Schwalbe. 2018. Artificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings? *BMJ global health* 3, 4 (2018), e000798.

[168] Ashley Marie Walker, Yaxing Yao, Christine Geeng, Roberto Hoyle, and Pamela Wisniewski. 2019. Moving beyond 'one size fits all': research considerations for working with vulnerable populations. *Interactions* 26, 6 (Oct. 2019), 34–39. https://doi.org/10.1145/3358904

[169] Jayne Wallace, John McCarthy, Peter C Wright, and Patrick Olivier. 2013. Making design probes work. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 3441–3450.

[170] Angeline Seng Lian Wan, S Mat Daud, Siao Hean Teh, Yao Mun Choo, and Fazila Mohamed Kutty. 2016. Management of neonatal jaundice in primary care. *Malaysian family physician: the official journal of the Academy of Family Physicians of Malaysia* 11, 2-3 (2016), 16.

[171] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.

[172] Katharina Weitz, Dominik Schiller, Ruben Schlagowski, Tobias Huber, and Elisabeth André. 2019. " Do you trust me?" Increasing user-trust by integrating virtual agents in explainable AI interaction design. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*. 7–9.

[173] Caroline E Wellbery. 2012. A case of medical uncertainty. *American Family Physician* 85, 5 (2012), 501–508.

[174] Caroline Whidden, Kassoum Kayentao, Jenny X Liu, Scott Lee, Youssouf Keita, Djoumé Diakité, Alexander Keita, Samba Diarra, Jacqueline Edwards, Amanda Yembrick, et al. 2018. Improving Community Health Worker performance by using a personalised feedback dashboard for supervision: a randomised controlled trial. *Journal of global health* 8, 2 (2018).

[175] Raynor William. 1999. The international dictionary of artificial intelligence.

[176] Susan Wyche. 2019. Using Cultural Probes In New Contexts: Exploring the Benefits of Probes in HCI4D/ICTD. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*. 423–427.

[177] Yao Xie, Melody Chen, David Kao, Ge Gao, and Xiang'Anthony' Chen. 2020. CheXplain: enabling physicians to explore and understand data-driven, AI-enabled medical imaging analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[178] Deepika Yadav, Anushka Bhandari, and Pushpendra Singh. 2019. LEAP: Scaffolding Collaborative Learning of Community Health Workers in India. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–27.

[179] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L Arendt. 2020. How do visual explanations foster end users' appropriate trust in machine learning?. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 189–201.

[180] Hugo Zylberajch, Piyawat Lertvittayakumjorn, and Francesca Toni. 2021. HILDIF: Interactive debugging of NLI models using influence functions. In *Proceedings of the First Workshop on Interactive Learning for Natural Language*

Processing. 1–6.