Public Health Calls for/with AI

An Ethnographic Perspective

AZRA ISMAIL*, Georgia Institute of Technology, USA
DIVY THAKKAR*, Google Research, India
NEHA MADHIWALLA, ARMMAN, India
NEHA KUMAR, Georgia Institute of Technology, USA

Artificial Intelligence (AI) based technologies are increasingly being integrated into public sector programs to help with decision-support and effective distribution of constrained resources. The field of Computer Supported Cooperative Work (CSCW) has begun to examine how the resultant sociotechnical systems may be designed appropriately when targeting underserved populations. We present an ethnographic study of a large-scale real-world integration of an AI system for resource allocation in a call-based maternal and child health program in India. Our findings uncover complexities around determining *who* benefits from the intervention, *how* the human-AI collaboration is managed, *when* intervention must take place in alignment with various priorities, and *why* the AI is sought, for what purpose. Our paper offers takeaways for human-centered AI integration in public health, drawing attention to the work done by the AI as actor, the work of configuring the human-AI partnership with multiple diverse stakeholders, and the work of aligning program goals for design and implementation through continual dialogue across stakeholders.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**.

Additional Key Words and Phrases: AI, ML, India, public health

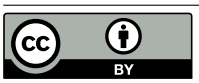
ACM Reference Format:
Azra Ismail, Divy Thakkar, Neha Madhiwalla, and Neha Kumar. 2023. Public Health Calls for/with AI: An Ethnographic Perspective. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 354 (October 2023), 26 pages. https://doi.org/10.1145/3610203

1 INTRODUCTION

Artificial Intelligence (AI) based systems are increasingly playing a role in decision-making and resource allocation in high-stakes settings, such as healthcare, public welfare, humanitarian crises, and more. The integration of AI in these contexts is frequently targeted towards supporting the efficient use of limited human and technical resources, and enabling more accurate and/or fairer decisions by stakeholders. Researchers in the field of Computer Supported Cooperative Work (CSCW) have been investigating how such systems might be designed appropriately [35, 37, 75], and have drawn attention to implications for historically underserved populations [9, 84]. Prior work has uncovered how AI can impact existing workflows, influence decision-making, and shape interactions across human actors [4, 65, 76]. This body of work also highlights risks of limited transparency, reduced accountability, and bias in AI systems, which are amplified in high-stakes

*Both authors contributed equally to this research.

Authors' addresses: Azra Ismail, Georgia Institute of Technology, Atlanta, USA, azraismail@gatech.edu; Divy Thakkar, dthakkar@google.com, Google Research, Bangalore, India; Neha Madhiwalla, nmadhiwala@gmail.com, ARMMAN, Mumbai, India; Neha Kumar, neha.kumar@gatech.edu, Georgia Institute of Technology, Atlanta, USA.



2573-0142/2023/10-ART354
https://doi.org/10.1145/3610203

settings [9, 63, 72]. As AI technologies make their way into public sector and health infrastructures, our work aims to inform these efforts by offering an ethnographic perspective on AI integration.

Prior research has highlighted the paucity of appropriately designed AI interventions targeting global health [37], calling for a more meaningful, human-centric integration of AI in the quest for societal impact. Our research offers a situated perspective of a real-world AI intervention that seeks to address this gap; we draw attention to key design decisions made when integrating AI into a complex maternal and child health ecosystem in a historically underserved context in Mumbai, India. Ensuring diverse perspectives, our research team includes multiple members who were involved in the design and deployment of the AI intervention we study, from established non-profit and industry organizations HealthNGO and TechOrg (anonymized), respectively. We focus on mCare, one of HealthNGO's largest programs, which delivers voice-based messages on pregnancy and child care to more than 240,000 beneficiaries (pregnant women and mothers) every year. A persistent challenge faced by mCare is the drop in engagement over the 18-month involvement with each target beneficiary, attributed to an array of factors influencing listening behaviors, including intermittent access to phones, cultural norms, health literacies, challenges experienced in the care journey, and more. To increase engagement in mCare, HealthNGO employs and trains human callers who encourage and offer counseling to beneficiaries through phone calls. Given the massive scale of this program, HealthNGO can only conduct a limited number of such calls. The introduction of the AI system in mCare is aimed at increasing overall engagement, by helping identify beneficiaries who may be most at risk of dropping out of the program and could benefit from human intervention.

Our paper presents an ethnographic investigation of a multi-stakeholder real-world AI-based public health intervention. We studied AI integration in the mCare program, conducting fieldwork across multiple sites in Mumbai between July to September 2022. Our research focuses on three sets of stakeholders—callers (including call center executives and hospital supervisors) at HealthNGO who were responsible for calling beneficiaries who were predicted to drop out of the program by the AI model, program and IT staff at HealthNGO shaping the design of the AI intervention and managing callers, and the development team at TechOrg building the AI model and supporting digital infrastructure. We began by observing the callers implementing the workflows on a daily basis, before turning to the perspectives of the program and development teams at HealthNGO and TechOrg respectively. The goal of our research is to inform public health projects that rely on data-driven approaches, by identifying critical considerations for the design of human-centered AI integrations. Our work seeks to answer the following research question: *How might a human-AI system be configured, implemented, and evaluated in support of multiple stakeholder perspectives in a large-scale public health setting?*

This paper presents one of the first ethnographic studies of an AI system deployed large-scale in a public health context, to the best of our knowledge. Our analysis details how different stakeholders attempted to determine the *what* or the program definition of the AI intervention, before uncovering complexities around determining *who* benefits, *how* the human-AI collaboration is managed, *when* calls must take place in alignment with various other priorities, and *why* the AI is sought, for what purpose. Our paper draws attention to the work done by the AI (as actor), the work of building the human-AI partnership (with multiple, diverse stakeholders), and the work of aligning program goals for design and implementation (through continual dialogue across stakeholders).

2 RELATED WORK

Our research connects with three bodies of work that have engaged in questions around the integration and design of AI/ML systems in the real-world. First, we inform research on AI in the public sector, by offering a multi-stakeholder perspective on the integration of AI in a large-scale

public health program in the Global South. We also build on a rich body of CSCW literature on design for Human-AI collaboration, that offers insight into the diverse human interactions that the development and implementation of such systems might entail. Finally, we contribute to emergent digital health literature at CSCW on the design and evaluation of AI systems.

2.1 AI in Public Sector Programs

There has been a growing interest, across research and practice, in the application of AI systems towards societal good. Many of these focus on how AI can assist decision-making, such as by supporting resource allocation at a systems level or providing worker support at an individual level. Fang et al. have previously developed an AI system to combat wildlife poaching, by serving as a decision aid to optimize the use of patrolling resources while handling complex topographic features and scheduling constraints [22]. Yadav et al. developed a decision-support system for staff at a drop-in center for homeless youth, to identify the most influential homeless youth to raise awareness about HIV (and other sexually transmitted infections) among their peers [95]. Several also target settings in the Global South, where limited data availability and varying data and digital literacies may pose additional challenges [72, 74, 86]. Among applications of such tools in healthcare, Killian et al. created a system to help community health workers identify which of their patients miss medications for tuberculosis frequently [49]. Another system by Nair et al. seeks to optimize health interventions that improve vaccination rates in Nigeria [60]. This growing body of work points to the uptake of AI among non-profit and public sector institutions to support work planning [59, 61], and the design considerations for such efforts [37]. We offer an ethnographic perspective to this space, to inform the design of AI systems that aim to support decision-making in large-scale public sector programs.

We also engage with critical perspectives on the use of AI tools in such settings, and attend to the harms that AI can perpetuate. Within CSCW, several scholars have examined the potential for harm with the integration of AI, for workers and other stakeholders. Brown et al. have examined how the general distrust in an existing child welfare service contributes significantly to low comfort in algorithmic decision-making [9]. Several other scholars have highlighted the conflicts that emerge between AI-assisted tools and workers' decision-making processes for caseworkers in child maltreatment screening and job placement programs [4, 46, 47]. CSCW research on algorithms in child welfare has also called attention to the power relationships across stakeholders, and emphasized the need for human-centered perspectives to algorithm design [75, 76]. In prior work on the development of a municipal decision-support systems for job placement, Møller et al. offer insights from how a team of data scientists, caseworkers, and system developers negotiated notions of value metrics and usefulness in a participatory design set-up [35]. The concerns highlighted by the research above has also prompted researchers to consider fairness and accountability in algorithms deployed in high-stakes public sector programs [87]. We add to this literature by drawing attention to the process of integrating AI on-the-ground in a public health context and the design decisions that it entails, determined through dialogue across diverse stakeholders.

2.2 Design for Human-AI Collaboration

The field of CSCW has begun to investigate how AI systems could be better designed to consider the needs of potential users as well as developers. Several industry actors have proposed guidelines for the design of human-AI interaction, targeting AI practitioners and designers [3, 24]. Researchers have studied how AI practitioners think about fairness, and considered the role of checklists and visualization tools in informing AI practice [34, 57, 73, 97]. Hohman et al. developed a design probe to understand how data scientists understand machine learning [32], and have also studied how data scientists wrangle with and iterate on datasets in machine learning [33, 74]. Chang et al. have

examined the potential role of collaborative crowdsourcing for labeling machine learning datasets [12]. Our research complicates such efforts by drawing attention to how AI development must be iterative and collaborative rather than relying on checklists, and how concepts such as fairness must be conceptualized within the context in which AI is deployed.

Several design methods and theories have also been developed to assist designers, such as for explainable AI [69, 93]. Dove et al. and Yang et al. have outlined the unique challenges in designing for machine learning-based systems and human-AI interaction [18, 98]. Grudin and Jacques similarly outline challenges when developing conversational agents, including when using human-bot hybrids [30]. Researchers have explored the evolving nature of Human-AI collaboration due to the introduction of AI systems into organizational workflows (e.g. [16, 62, 70, 92, 99, 101].) Oh et al. conducted a user study on a co-creation AI tool, indicating user preferences for explanation [62]. Suh et al. examined the process of co-creating music with generative models and discovered the unique role of AI as a social glue in enabling collaboration between musicians [81]. Prior work by Khadpe et al. has examined how humans view the competence of conversational agents and the moments they would like to be interrupted by AI agents in their workflows [48]. Wang highlights the concerns of data scientists while collaborating with AutoML technologies [92]. Gal et al. developed new workflows to improve human-AI collaboration where humans aid the machine in solving difficult tasks with high information value and the machine can generate motivational messages that highlight different aspects of collaboration [25]. In their survey paper on human-AI collaboration, Lai et al. call for a need for common frameworks to account for a range of research and design spaces [52]. Prior work has also brought attention to the value of moving beyond algorithmic interventions, to increase explainability for improved human-AI collaboration [19]. We contribute to this growing body of work by examining how AI plays a role as an actor in a public health context, and the configuration that human-AI partnership might entail.

Several studies have begun to also investigate the role of values in Human-AI collaboration. A study on the perceptions of algorithms used by Wikipedia, uncovered the need for these to be transparent and align with community values, and allow human actors to act as the final authority [79]. Prior work has also examined the role of values in AI systems and datasets [39, 77, 86], foregrounding the interplay of social, cultural and organizational factors that impact AI systems. There have also been experimental studies on how users perceive AI trustworthiness or credibility (e.g. [5, 40, 41, 66, 67]), or attempt to "trick" AI systems to protect their data privacy (e.g. [88]). We take inspiration from this body of work when analyzing an instance of AI integration that prioritizes the agency of workers. Researchers have examined the tradeoffs between team performance and model accuracy and found that optimizing for model accuracy is not sufficient to improve team performance, even in high-stakes settings [5, 6]. Ehsan et al.'s work also points to how the effects of algorithms can persist much longer after the algorithm is removed [20]. Our paper speaks to the need to think about the role that might AI play when introduced into a context, and carefully define success metrics for AI interventions, given their potential lasting effects.

2.3 Human-AI Collaboration in Healthcare

CSCW researchers have also begun to design and evaluate AI systems in healthcare. There have been a small number of observational and participatory design studies that aim to inform the design of systems and datasets in this area (e.g. [21, 38, 55, 71, 100]). A rare example of a study of the real-world deployment of an AI system is Beede et al.'s ethnographic study of a deep learning system for diabetic retinopathy in hospital settings [8]. Given that there are limited opportunities for design iterations in the case of AI systems post-development, they argue for "formative research that provides a strong understanding of clinical users and their context is critically important to the success of such a system" [8]. Cai et al. have studied how AI could support routine workflows by

introducing tools and interfaces that would help clinicians work with AI decision-making systems to meet their unique needs of examining ML outputs in a clinical setting [10]. Others have examined AI for decision-making with healthcare providers in rural clinics and rehabilitation assessment [54, 91]. We extend this work to look at decision-making in a public health context.

Perspectives from and/or situated in the Global South are also increasingly seeing representation in this discourse. Yadav et al. have studied the potential of chatbots for breastfeeding education by conducting a Wizard-of-Oz experiment with FHWs and breastfeeding mothers [96]. Okolo et al. have also studied the perceptions of health workers on the introduction of AI in their daily workflows [64]. We align with this emerging work, as well as in a rich body of prior literature in the field of Human-Computer Interaction for Development (HCI4D) that has detailed considerations for technologies integrated in diverse settings, particularly when working with underserved populations [17]. This includes extensive documentation on the gendered and inter-mediated access to technology [36, 50, 80], differing notions of privacy [2, 44], and more. Our research builds on this history, to offer an understanding of how AI systems might be integrated in public health, while considering the sociocultural context within which it is embedded.

3 BACKGROUND

Our research takes place in the context of HealthNGO's mCare public health program and TechOrg's AI intervention (anonymized). Below we share an overview of both stakeholders and their involvement, critical to understanding the contributions of our research.

3.1 Overview of the mCare Program

mCare is a free mobile call service operated by a maternal and child health non-profit organization called HealthNGO, headquartered in Mumbai, Maharashtra (India). The target beneficiaries of the program are pregnant women and new mothers. Women are enrolled—with written consent—by a health worker during a home or hospital visit. As part of the program, they receive timed recorded voice messages every week, corresponding to their gestational age or the age of their child (till they are one year-old). The voice messages provide information on breastfeeding, sanitary practices, nutrition, child development, and more. If the woman misses the call at the scheduled time, she receives a call the following day at the same time, and a third time on the day after that if she misses again. If the woman would like to listen at a time of her choice, she can call a number to hear the voice message for that week. mCare has reached over 2.6 million women across nine states in India as of 2022, and offers content in four languages. A recent evaluation of the mCare program through a three-year randomized control trial demonstrated that the calls have had a positive impact on infant birth weight, infant feeding practices, and immunization.

HealthNGO has set up a call center in their office in Mumbai where they run several programs to support mCare, mostly focusing on beneficiaries in the state of Maharashtra. The calls are placed by women Call Center Executives (CCEs). One of their largest programs is the *37-week program*, where *service calls* are placed by CCEs around the 37th week of pregnancy to determine if and when women have delivered their baby, based on which they activate the calls on child care. Once a month, HealthNGO sends a text to inform mothers that they can *opt out* of the program over a phone call, particularly in the case of miscarriage or child death. Finally, they also have a *missed call service* that beneficiaries can call to ask questions around maternal or child health, to report a delivery and activate the calls for child care, or to report a miscarriage or child death and stop the calls. All programs are free for beneficiaries.

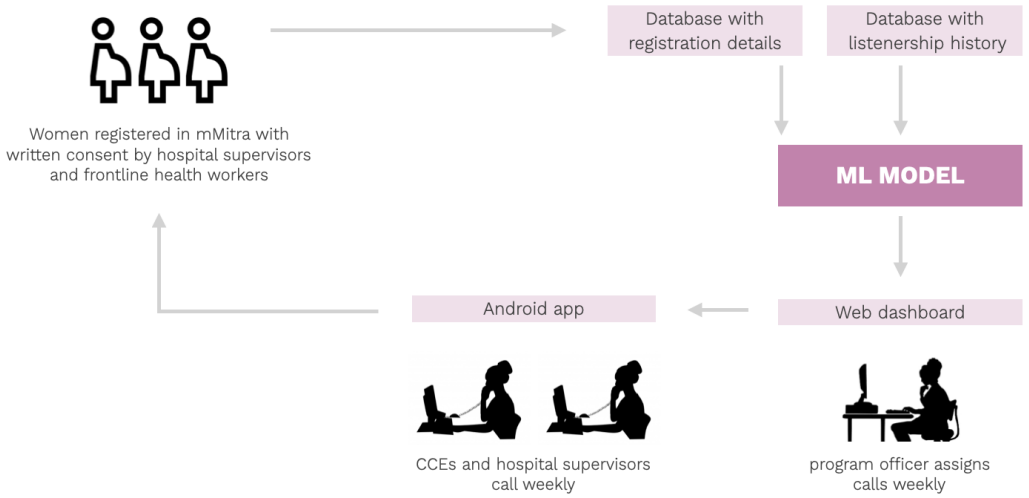


Fig. 1. **Program flow:** Beneficiaries are registered by hospital supervisors and community health workers. Their registration data after anonymization (including demographic information, preferred language and time of the call) as well as listenership history is used to predict beneficiaries likely to drop out with the machine learning (ML) model. The list of beneficiaries generated is uploaded to a web dashboard by TechOrg where it is viewed by a program officer at HealthNGO, who then distributes it across hospital supervisors and CCEs. The CCEs and hospital supervisors receive a list of beneficiaries to call every week on an Android application.

3.2 Integration of AI into the mCare Program

AI was introduced into the mCare program with the goal of *predicting beneficiaries who were likely to drop out of the program.* Given the massive quantities and real-time nature of data, identifying beneficiaries manually was challenging. Also, due to limited human resources, only a certain number of live calls were possible every week. In 2019, HealthNGO began a partnership with TechOrg towards developing a machine learning (ML) system that could automate the process of identifying beneficiaries to provide a follow-up call. At the start of the collaboration, HealthNGO tested alternatives to calling, including text messages, and text messages followed by calls to only those beneficiaries whose engagement had not increased. They found live calls to have the greatest impact and decided to focus on this approach despite it being the most expensive and labor-intensive option. Through a collaborative and iterative process over two years, TechOrg developed, deployed, and evaluated several ML models. Figure 1 details the overall workflow surrounding the ML model.

3.2.1 Development and Evaluation of the ML model. The current ML model in deployment uses the Restless Multi-Armed Bandit (RMAB) framework, which has been used to tackle resource allocation problems in other domains [42]. The model is trained on historical call log data as well as socio-demographic features collected during enrollment in mCare such as age, language, income range, and gestational age. It is designed to *maximize engagement* by selecting beneficiaries who are at risk of dropping-off from the automated calling service and could benefit from a live call (AI-assisted intervention) to stay engaged. For the model, engagement was defined by HealthNGO as listening to at least one call in a week for at least 30 seconds. The model accounts for the possibility of transitioning from engaging to non-engaging state. TechOrg has been working closely with HealthNGO to define key features such as the frequency of repeated live service calls, number of live service calls, and metrics for engagement. The ML model was first tested with 23,000

participants through a randomized controlled experiment, and showed an increase in listenership of 30% compared to the current standard of care group. The ML model has been revised since, and the evaluation of beneficiary engagement with the new model has been ongoing for several months as of September 2022.

3.2.2 Workflows Associated with the ML Intervention. The ML intervention has been deployed in mCare for over a year, and is being used by TechOrg *to generate a list of beneficiaries for callers at HealthNGO to reach out to every week*. Once research staff at HealthNGO receive the list, they use a web application to allocate calls to various callers, who receive them on a mobile application. Both applications have been developed by the IT team at HealthNGO, with TechOrg's support. Beneficiaries who have been registered at government hospitals or partner NGOs are assigned to their respective hospital supervisors or NGO staff, and the remaining are randomly assigned to CCEs. We use the term *callers* to refer to both CCEs and hospital supervisors. The decision to also have hospital supervisors make service calls was for two reasons: CCEs were already busy with existing programs and more human resources were needed, and HealthNGO wanted to leverage the close field interactions that hospital supervisors had with beneficiaries. Callers receive the list of the beneficiaries on the mobile application at the beginning of every week, and have to make three attempts to reach them that week. If an attempt is not "successful" because no one picks up the call, then they call again the next day. If no one picks up after three attempts, then the call is marked as "unsuccessful" on the app. After every call, callers complete a short survey on the outcome. The *call success rate*—or the percentage of beneficiaries that picked up the call—as well as the *call outcomes* are monitored by HealthNGO.

4 METHODS

Our research objective was to gain a deep understanding of the integration of AI into broader care ecologies. To this end, we conducted a multi-sited ethnographic study [58] of a machine learning system deployed as part of the mCare program in Mumbai (India) over six weeks in July-September 2022. Drawing inspiration from Marcus' recommendations, we *follow the AI* (our object of study), across multiple sites and stakeholders [58]. Our research took place across three types of sites: HealthNGO's call center (which also operated as their office), government hospitals, and home visits to beneficiaries enrolled in the program. We conducted observations, interviews, and focus group discussions, as well as content analysis of materials that had been generated over the history of the project, including documentation and study protocols. The research was approved by the Institutional Review Board at an Anonymous US University and due permissions were granted by the Indian institutions that were participant to this research.

4.1 Participant Information

We investigated AI integration from the perspectives of those implementing and using the AI system, including callers (CCEs and hospital supervisors), program staff at HealthNGO, and the development team at TechOrg. As we detail in the section on limitations and future work, we chose not to directly engage with beneficiaries because of early interactions with HealthNGO that revealed that on account of their decision to explicitly integrate AI in the background of their public health program, beneficiaries only interfaced with human actors over a phone call and had no interactions with the AI system. We conducted a total of 24 interviews, 2 focus group discussions, and approximately 90 hours of observation. Our research participants included 32 stakeholders at HealthNGO and 4 at TechOrg, at multiple levels of the organization. We conducted interviews and observation with all three Call Center Executives (CCEs) and nine hospital supervisors who were conducting service calls for the AI intervention. In addition, we observed nine CCEs as they

Pseudonym	Title at Organization	Gender	Highest Education Attained	Years in the Role	Field Site	Participation
Kusum	Staff (HealthNGO)	F	High School	7	Office/Call Center	I + O
Shanti	Staff (HealthNGO)	F	High School	6	Office/Call Center	I + O
Parvati	Staff (HealthNGO)	F	High School	5	Office/Call Center	I + O
Geeta	Staff (HealthNGO)	F	-	-	Office/Call Center	O
Neelam	Staff (HealthNGO)	F	-	-	Office/Call Center	O
Lata	Staff (HealthNGO)	F	-	-	Office/Call Center	O
Radhika	Staff (HealthNGO)	F	-	-	Office/Call Center	O
Khushboo	Staff (HealthNGO)	F	-	-	Office/Call Center	O
Vanita	Staff (HealthNGO)	F	-	-	Office/Call Center	O
Riya	Staff (HealthNGO)	F	-	-	Office/Call Center	O
Monica	Staff (HealthNGO)	F	-	-	Office/Call Center	O
Babita	Staff (HealthNGO)	F	-	-	Office/Call Center	O
Seema	Middle Management (HealthNGO)	F	High School	6	Office/Call Center	I + O + FDG1
Meena	Middle Management (HealthNGO)	F	High School	7	Hospital 1	I + O
Reshma	Middle Management (HealthNGO)	F	High School	7	Hospital 2	I + O
Archana	Middle Management (HealthNGO)	F	High School	7	Hospital 2	I + O
Meghna	Middle Management (HealthNGO)	F	Undergraduate Degree	6	Hospital 3	I + O
Sunita	Middle Management (HealthNGO)	F	Undergraduate Degree	8	Hospital 3	I + O
Jyoti	Middle Management (HealthNGO)	F	High School	7	Hospital 4	I + O
Fatima	Middle Management (HealthNGO)	F	High School	7	Hospital 5	I + O
Rani	Middle Management (HealthNGO)	F	High School	7	Hospital 6	I + O
Leena	Middle Management (HealthNGO)	F	High School	8	Hospital 6	I + O
Poonam	Middle Management (HealthNGO)	F	High School	8 months	Home Visit	I + O
Deepa	Middle Management (HealthNGO)	F	Undergraduate Degree	6 months	Home Visit	I + O
Amrita	Middle Management (HealthNGO)	F	Undergraduate Degree	6	Hospital 1	I + O
Sonam	Research Staff (HealthNGO)	F	Graduate Degree	5	Office/Call Center	I + O + FDG1
Nita	Middle Management (HealthNGO)	F	Graduate Degree	9	Office/Call Center	I
Preeti	Middle Management (HealthNGO)	F	Undergraduate Degree	7	Office/Call Center	I
Ambika	Middle Management (HealthNGO)	F	Graduate Degree	4	Office/Call Center	O + FGD1
Mohit	Middle Management (HealthNGO)	F	Undergraduate Degree	7	Virtual	I
Paresh	Technical Staff (HealthNGO)	M	Graduate Degree	-	Office/Call Center	I + O
Gavin	Middle Management (HealthNGO)	M	Graduate Degree	7	Virtual	I
Shruti	Technical Staff(TechOrg)	F	-	-	Virtual	FGD2
Karthik	Technical Staff(TechOrg)	M	-	-	Virtual	FGD2
Disha	Technical Staff(TechOrg)	F	-	-	Office/Call Center + Virtual + Home Visit	I + O + FGD1 + FGD2
Chris	Technical Staff(TechOrg)	M	-	-	Office/Call Center + Home Visit	I + O + FGD1

Table 1. Demographic information about our study participants. "Staff" refers to the title of Call Center Executive. Our study engagement included interviews (I), observations (O), and focus group discussion (FGD1 was in-person while FGD2 was virtual). "-" means that the data was not collected.

conducted service calls for the 37-week and missed call programs, and two field investigators conducting home visits to do surveys. We also interviewed HealthNGO's program officers and members of their IT and research teams. Finally, we conducted an in-person focus group discussion with three program officers at HealthNGO and two AI developers, and a virtual focus group discussion with four AI developers at TechOrg. Table 1 provides more details about our study participants and the nature of their engagement in the study. To preserve anonymity, we have not listed the specific affiliations of the participants on the program and development teams at HealthNGO and TechOrg respectively.

4.2 Data Collection

Given the ethnographic nature of our engagement, our data collection involved a combination of extensive handwritten field notes, audio recordings, and photographs and videos. Our interviews lasted between fifteen to sixty minutes, and were audio recorded with consent. We took handwritten notes during all our observations at the call centers, hospitals, and home visits. At times, we asked clarifying questions or engaged in short conversations with study participants during observations, taking care to ensure that their work was not disrupted. All data collected was anoymized, and photographs and videos were taken with informed consent.

4.2.1 Field Site 1: HealthNGO Call Center and Office. We conducted observations at the call center (which is housed within the office in a dedicated room), as CCEs placed calls using the mobile application. To better understand how the AI program fit into the broader ecosystem, we also observed the other CCEs who were conducting calls as part of the 37-week and missed call programs

(described earlier). The calls we observed took place in Hindi or Marathi. During these calls, we could only hear the CCEs' responses as they used earphones or a corded telephone, preserving the privacy of the beneficiaries. We took great care to ensure that no identifiable information about beneficiaries was recorded. We also interviewed other actors at HealthNGO associated with the program to understand the history of the program, their role, experience with the program, and challenges encountered. These interviews either took place in HealthNGO's office, or over a phone call or Zoom depending on the availability of the participant. We also observed these actors in their everyday workflows and interactions relating to the mCare project at the HealthNGO office.

4.2.2 Field Site 2: Government Hospitals. Once we reached saturation in data with CCEs, we began observations and interviews with hospital supervisors who were also conducting calls as part of AI program, in addition to their role in registering beneficiaries for mCare. Hospital supervisors typically had a desk located near the antenatal and postnatal care unit in the government hospital where they were placed by HealthNGO. We observed how hospital supervisors registered new beneficiaries for the mCare service, their service calls with beneficiaries using the same mobile application used by CCEs, and in-person follow-ups with registered beneficiaries about mCare.

4.2.3 Field Site 3: Beneficiary Home Visits. We accompanied two field supervisors during their home visits to conduct surveys with three beneficiaries who were enrolled in the mCare program. These helped us get a broader understanding of the contexts that beneficiaries were coming from and their experiences with the mCare service. During these visits, the field investigators completed a survey to gather feedback about mCare.

4.2.4 Focus Group Discussions. We engaged in a two-hour long focus group discussion (FGD1) with two members of TechOrg and three program officers at HealthNGO's office. We also conducted an hour-long virtual focus group discussion (FGD2) on Google Meet with four developers at TechOrg. The goal of the FGDs was to facilitate discussion on the challenges encountered when translating health program goals to an AI intervention, and to reflect on the the implications of the findings from our fieldwork for the design of the AI intervention, particularly in relation to model performance and fairness.

4.3 Data Analysis

Data was collected by the first author in Hindi, Marathi, and English. All data was translated and transcribed by them to English. Since Marathi is not the first author's native language, at times they relied on the study participants to help translate into Hindi or English. We analyzed our transcribed audio-recorded interview data, written observation notes, and other documents through an iterative inductive coding and memo-writing process, as recommended by Charmaz [13]. During the process of collecting and coding our data, we also engaged in memo-writing to help reflect on themes emerging from our data while doing fieldwork. The memos were discussed among the authors on a weekly basis during fieldwork, to in turn inform data collection until we reached saturation in data for a particular theme [13]. We conducted several rounds of coding, with the first round resulting in sentence-level codes which could be a result of either an observation or interview. For instance, codes such as "the call was not picked up" and "the phone is generally with the husband" were a result of observations of calls by CCEs. Codes such as "the goal of AI is to increase beneficiary engagement" or "should be a way to get beneficiary feedback" were a result of interview data. In the second round, we began to group together codes based on common themes emerging from the data, combining sentence-level codes from either interview or observation data to arrive at a broader understanding of an interaction or workflow. For example, codes included "inferring religion based on beneficiary name", "changing the timing of call", and "providing counseling to a woman who

miscarried". The third round of coding finally led to the generation of high-level themes such as "identifying miscarriages with AI", "determining the timing of the calls", and "problem-solving on the calls". In each round, the codes were discussed among the authors at regular intervals as they were refined until consensus was achieved.

4.4 Positionality

Authors on this paper represent diverse perspectives from HCI, AI development, and public health, and have extensive research experience around technology design for maternal and child health programs in India. Our team includes members from HealthNGO and TechOrg who have been closely involved in developing and implementing the AI program for more than two years; their perspectives were crucial for gaining access to the research sites and participants, framing the research direction, and reflecting on the relevance of the insights for future AI development and integration. All four authors on this paper are of Indian origin; two are based in India and the other two routinely cross borders between India and the US.

5 FINDINGS

We now present findings from our ethnographic fieldwork, combining what we learned through interviews and focus groups with various stakeholders and what we observed while the intervention was in play. Throughout, our focus remains on the details of integrating AI into a complex public health ecosystem, and considerations around the configuration, implementation, and evaluation of the system. Specifically, we focused on who the AI is directed at, how and when it intervenes, and why/whether it brings value, as we detail below.

5.1 Who to Target: Identifying Beneficiaries

Who benefits from the AI intervention was a key question for our investigation, as is likely to be the case for most societal impact interventions like the one we studied. Our findings highlight that design decisions regarding who should be targeted are not always straightforward ones to make, for numerous reasons. How marginalized target beneficiaries are (and how marginalization is assessed to begin with), how equipped they are to partake of the intervention, and how great their need is for the intervention—all play a crucial role in determining the efficacy of the ML model at play, as we present below.

5.1.1 Identifying the Marginalized. In the demographic context where our research is situated, historical marginalization based on identities like caste and religion have been recognized to have significant impact on access to health services [7, 31, 68]. Program staff and hospital supervisors explained that collecting such data could leave participants vulnerable to data misuse, create hesitation in joining the program, and might not be appropriate in a setting like a hospital where others could overhear the discussion thus leading to involuntary disclosure of their identity. It was also preferred that the callers not have this information so that they treated each beneficiary without bias. Details visible included only name, age, date of last menstrual period, gestational age, number of call attempts made, and the outcome of the calls. It is worth noting, however, that caste and religion details are frequently determined based on last name; when we asked what language callers spoke to beneficiaries in and how they decided, Kusum (CCE) shared: *"We decide based on the name. If the name is Muslim then we speak in Hindi, otherwise we speak in Marathi. We can tell."*

At present, the ML model in use does not look at caste or religion, but uses income and education as markers of marginalization instead. Income information, however, is not collected by most hospital supervisors during the registration process, we were told (and also observed). Instead, these supervisors ask beneficiaries for their husband's occupation, and attempt an approximation

of the income accordingly. Sometimes hospital supervisors did not ask about the occupation either, but put down their own perception based on their interaction with the beneficiary or their family, or based on other demographic details collected such as education level and the type and ownership of the phone. To address this inconsistency, HealthNGO has recently redesigned their registration form to replace income with the husband's occupation. Prior work by Suresh et al. has reflected that annotating for sensitive demographic information, such as race, can be ethically fraught, and emphasized the need to stay with such tensions in AI development work [85]. This is also what we observed of our participants; it remains a program goal to effectively categorize beneficiaries based on backgrounds to identify the most marginalized, but it is not one that is straightforward to implement due to social and cultural factors that are outlined above.

5.1.2 Identifying Potential for Impact. As we explained in the background section above, AI was intended to increase engagement from beneficiaries who were perceived to be "low listeners" or less engaged in the program. A significant challenge with selecting the beneficiaries who were least engaged was that call success rates were very low. Our interviews revealed that low levels of engagement could be attributable to a number of reasons, such as connectivity issues, lack of mobile credit to receive calls, the phone number having changed, or the phone being with the husband or a family member. These types of challenges were also more likely to be experienced by women from marginalized backgrounds who might not have their own personal phone or reliable network access. There was agreement that it was critical for the program to reach these beneficiaries. Focusing on low listeners, however, came into conflict with the experience for callers, who experienced a drop in their motivation levels if their calls were not answered. For instance, during one of our observations, a CCE (Parvati) expressed her frustration at the rate at which her calls were going unanswered: *"Today, none of my calls are connecting! How are others' calls connecting?"*

This conflict surfaces a tradeoff that must be made between the efficacy of the program via targeting lesser engaged beneficiaries and the sustainability of the program via ensuring that the callers are motivated to execute on their responsibilities. To achieve a compromise, the current ML model leaves out beneficiaries who have never answered a single call, with the assumption that the number is likely not operational. Our focus group discussion with TechOrg stakeholders revealed that they were considering using an autocaller to determine if a number was active and ringing before sending it to a caller.

5.1.3 Identifying Need. HealthNGO made a significant effort towards identifying women who were not in need of the information being provided. In some cases this meant excluding women who were simply not interested in the program. For example, Fatima (Hospital Supervisor) shared, *"There are some who are more educated or this is their second or third child so they feel like they already know everything and don't need to listen."* Reshma (Hospital Supervisor) also pointed out that if the beneficiary was more educated, they could find similar information on the internet such as on YouTube, and might find videos more appealing than voice calls.

HealthNGO was also interested in determining who was no longer engaged in the program, or numbers that were no longer active or repeatedly out of coverage. For example, Kusum (CCE) reported: *"The husband requested to switch off the calls because of his job. He is a driver and travels a lot, and was saying he gets very disturbed by the repeated calls".* The motivation was to stop calls to these numbers, towards conserving resources that were already acutely constrained (time and money). Prioritizing informed consent however, HealthNGO did not switch off calls to anyone who had not actively opted out. The ethics around taking such an action was a topic of discussion across various stakeholders in the program. A major concern was around identifying false positives, in case the woman was interested in the program but struggling to access it due to systemic barriers.

On the other hand, program officers also had ethical concerns with calls that were currently going to women who had miscarried, had an abortion, or lost their child. Many beneficiaries had provided initial written consent to opt into the mCare program but struggled to opt out later due to limited digital literacies. Our interviews revealed that the use of AI was particularly helpful in identifying and actively reaching out to such women. The caller could then provide emotional support, and also immediately switch off calls to them with their consent to prevent potential trauma. A hospital supervisor emphasized the value of such interactions:

> "The lady's husband picked up the call, and I started telling him about the mCare service, and he said—'Madam, I want to stop the calls.' When I asked him why, he said that his wife had passed away while delivering their child. So he was very upset. But our calls had still been going to him because it was his number... Our job is just to listen. He was even crying, and there wasn't much that I could have said but a rapport was built."—Leena, Hospital Supervisor

Our interviews revealed that beneficiaries had to call a specific number to opt out of mCare. They then received a call back from HealthNGO to get their verbal consent to opt out. The number to call was shared during enrollment and monthly text reminders were also provided, but limited digital literacies could mean that many were unable to avail the service.

5.2 AI in the Background: Interactions between Callers and Beneficiaries

The goals of the AI intervention are achieved, or not, based on the day-to-day interactions and the sustained relationships between the callers and beneficiaries. We focus now on the "humans in the loop" or the callers and how they contribute to the efficacy of the AI system—by enrolling and informing beneficiaries, sustaining their engagement, hustling to connect with beneficiaries as needed, and gathering feedback for improving the system's performance.

5.2.1 Enrolling and Informing Beneficiaries. Our interviews with callers confirmed that their main goal was to increase engagement among low listeners, in line with the goals of the ML intervention. Kusum (CCE) shared, *"For the ML program, they told us that those beneficiaries that are listening less, we need to call them and convince them to listen to the calls."* A necessary outcome for beneficiaries to be engaged was awareness about mCare, starting with the hospital visit when they were enrolled in the program. Though most beneficiaries were aware, several had forgotten that they had enrolled or thought they were receiving spam calls. Our observations revealed that hospitals in particular were stressful environments for many beneficiaries, who did not always retain the information that they were being enrolled into the mCare program. As a hospital supervisor mentioned, *"Women don't remember that they registered by the time they go home. A lot is happening during their visit. When they get a call, then they remember."* Brochures with information about mCare were distributed at the time of enrollment, but this was not a reliable mode of communicating mCare's goals, since the brochures could be easily misplaced; it was also possible that the women were not textually literate. A key first step to driving engagement with the ML intervention was therefore to ensure that the beneficiaries were appropriately informed when enrolled.

5.2.2 Sustaining Beneficiary Engagement. The calls were key for strengthening the relationship between the HealthNGO's callers and beneficiaries. One CCE described how it was important to *"build rapport"* on the call. We found that it was relatively harder for CCEs to build rapport than it was for hospital supervisors who were assigned to beneficiaries at the time of registration at their hospital and had likely met them in person. In the calls made by hospital supervisors, they reminded the beneficiaries of their visit, and the repeat interactions helped build trust and *"become kind of a friend."* One hospital supervisor also shared that *"the more time that a hospital supervisor*

spends with the beneficiary, the more questions come up," alluding to the richer information exchange that took place between them and the beneficiaries, as compared to that with the CCEs. We found that calls by CCEs were typically shorter; they were also assigned more calls and did not have experience with repeated/in-person interactions with beneficiaries.

Remote interactions in turn shaped in-person interactions with hospital supervisors, who shared that beneficiaries would meet them on their next visit to the hospital and mention that they had received a call from them. Fatima described this also created a sense of accountability by creating the perception across other beneficiaries who were there in person that *"the hospital will know if I don't listen."* Apart from follow-up on calls, hospital supervisors were also responsible for follow-up with registered beneficiaries in person during their hospital visits to understand if they were facing any challenges with the mCare service. During the calls and in-person interactions, beneficiaries also sometimes asked health-related questions as well and received counseling.

The above data highlights that the nature of relationship that a beneficiary could nurture with a CCE was quite different from that with a hospital supervisor, but the latter was not part of the ML intervention, and any improvements in engaging beneficiaries that resulted from interacting with the supervisor would therefore not be factored into assessing the efficacy of the ML system.

5.2.3 Hustling to Connect. Our observations revealed that a key aspect of the call was not just informing the caller about the program, but uncovering why they were not engaging and addressing this lack of engagement. A significant challenge highlighted by callers was that *"most of the time, the phone is with the husband"* who was not at home when the calls were scheduled. According to Seema (Call Center Team Lead), roughly forty percent of the calls were picked up by the husband. In such cases, the callers tried to determine when he was likely to be at home to change the timing of the call accordingly. If this was not possible due to the nature of their job, for instance, if they had uncertain hours as a driver, they would ask them to record the calls. Several husbands that callers interacted with were already recording the call, and later played them to their partners. If the husbands did not know how to record, then they encouraged them to be at home (with their wives) at the time of the call. In some cases, women had provided their husband's number due to privacy concerns or had since gotten their own phone. Callers asked to switch the phone number such situations.

There were other reasons for lack of engagement that the callers tried to address. For instance, if the woman had delivered the baby and was still receiving pregnancy calls, they updated the delivery date. In some cases, the date of the last menstrual period had been recorded incorrectly or the woman had miscarried and then became pregnant again, so the date had to be updated so that the information aligned with their stage of pregnancy. In the case of a miscarriage, abortion, or infant death, callers typically kept the call short and apologized for calling, and switched off the service. In a few cases that we observed, callers asked the women if they were trying to get pregnant again, and advised them to wait for a few months and take care of their health. The movement of beneficiaries also shaped their engagement with the program. It was a cultural practice in this context for women to move to their parents' home around their ninth month to give birth, and then move back a few weeks after they had given birth. During this period, many were inaccessible. If they were using their husbands' phones, they no longer had access. If they were using their own phones, some did not take it with them due to mobile roaming costs. In such cases, the caller would ask for an alternative number that was accessible to the woman. The calling patterns of beneficiaries may have been different in each of these cases. To take into account these varied scenarios, developers at TechOrg shared during our focus groups that they would be adding call outcome as a feature in the next iteration of the AI model.

5.2.4 Getting Feedback about the Program. Our interviews with callers revealed that they enjoyed interacting with beneficiaries, and were particularly motivated by positive experiences related by beneficiaries. They also saw the calls as a way to get feedback on the mCare program. For instance, several callers shared that a common complaint by beneficiaries was that they were not receiving calls, which has implications for the behavior presumed by the ML model. Our interviews with callers revealed that this was likely due to network issues or because the mobile balance had run out (some balance is required to receive calls, even if the service is free). Hospital supervisors, in particular, were required to share case studies from their beneficiary interactions, whether in-person or on the call. They found that the calls served as a channel to get feedback and stories that might not otherwise come up in their in-person follow-ups, such as around miscarriages, movement patterns, and husbands' perspectives.

5.3 When to Call: Aligning Across Constraints

Our findings highlighted that key program decisions needed to be made around *when* calls were placed. The response rate on calls was also an important consideration from the perspective of designing the AI system, since calls took up time—a precious resource, and it was crucial to protect against failed call attempts. Below we discuss how calling needed to be brought in alignment with the care journey, availability, workflows, and adjacent programs.

5.3.1 Aligning with the Care Journey. Our discussions at HealthNGO revealed that the mCare program regularly saw a linear drop in engagement in the first few months after registration, and then engagement became fairly stable. There was a second significant drop after giving birth, before engagement became stable again. The second dip was also associated with movement patterns mentioned earlier, as many women moved to their parents' homes during the last month of their pregnancy. Given these complexities, one of the aspects of program design that was discussed by HealthNGO and TechOrg was the "best time" to place calls. Currently, the calls were placed within three months of an individual's registration in the mCare service, to prevent the initial drop in engagement. As most women were registered in their third to fifth month of pregnancy, the calls largely went to women who were pregnant. By calling earlier in the program, HealthNGO was also able to identify miscarriages earlier and prevent potential repeated trauma. Prior work by Chen et al. has discussed how computing systems could be trauma-informed, and the approach followed here aligns with this perspective [14].

5.3.2 Aligning with Availability. Our observations and interviews with callers highlighted the importance of getting the timing of calls right. Given that phones were often shared within the household, and that the phone owner was likely to be away from home (i.e., not with the beneficiary), there were recurring discussions with stakeholders around whether and how to schedule the calls. Frequently during registration, we observed that beneficiaries were unsure about when they would be available to receive calls or have access to the phone. The hospital supervisor would suggest times based on who owned the phone and perceptions about how women in this context organized their day. If the husband owned the phone, then they typically suggested an evening slot. We found that though callers had differing perspectives on when to call, evening was considered a good time to call, in general. Hospital supervisors typically called from home in the evenings after they had completed their duties at the hospital, but CCEs could not typically talk then due to limited working hours. Also, since hospital supervisors had fewer people to call, they could more easily try to vary times for greater success in successive attempts. This was harder for CCEs, possibly due to fixed hours and the volume of calls they made every week that made it difficult to keep track. Our interview with Sonam confirmed that hospital supervisors typically had higher call success rates than CCEs.

5.3.3 Aligning with Workflows. We found that workflows that were ideal for beneficiaries sometimes conflicted with caller workflows. Callers were only available to conduct calls during certain times of the day. CCEs had fixed working times at the office, and hospital supervisors had other responsibilities and were available based on the working hours of their respective hospital. Even if a hospital supervisor or CCE was willing to take calls outside working hours, there were other concerns. For instance, Rani (Hospital Supervisor) suggested that there be an option for the beneficiary to call them back. She shared:

> *"See it's like this, many times a child picks up the call and the mother is not available because she is sleeping. So we tell them that we will call again at a different time. But because the call is from the hospital, they can get worried about why they got a call. There should be a way for them to call back to check, and we have also already given time."*
> —Rani (Hospital Supervisor)

By design, an option for beneficiaries to call back was not currently available. During the pandemic, hospital supervisors had conducted calls of a similar nature from their homes. Those were from their personal numbers, and resulted in beneficiaries calling them back on their personal number at various times of the day or even harassing them. Learning from that experience, the calls in the AI program were conducted through a centralized number to protect the personal numbers of the callers. This was relatively more expensive, since two calls needed to be made, from the caller to the central number, and then to the beneficiary. HealthNGO had to determine priorities in such cases, which was to ensure that caller privacy was maintained, despite additional monetary cost.

5.3.4 Aligning Across Programs. Our fieldwork revealed that the long-term implementation of the AI intervention had also stimulated broader questions among the staff at HealthNGO and developers at TechOrg on how their programs could be more effective overall. Currently, the calls placed based on the AI's recommendations were considered a separate program by HealthNGO, with a dedicated pool of callers. Our observations of callers, however, conveyed that the calls conducted were very similar to those in the 37-week program, though the latter was not specifically geared towards increasing engagement. The 37-week program was aimed at determining when women had delivered their baby to start the automated voice messages on child care. If the woman was still pregnant, then the callers noted her expected date of delivery. One possibility that we observed being discussed across organizations was to overlap these programs and potentially implement two rounds of calling with AI, because the 37-week call had too narrow an objective and could be better aligned with the goal of increasing engagement overall. The first round would focus on miscarriages and the initial dip in engagement, while the second would address movement patterns and the later dip in engagement.

5.4 Why AI: Considering Program Goals

Our interactions with diverse stakeholders revealed that there were multiple reasons why AI was considered to be of value. Below we list key considerations around AI-driven outcomes for mCare.

5.4.1 Increasing Program Engagement. Increasing engagement of target beneficiaries with mCare was the goal of the AI intervention, which was developed to predict callers whose engagement was likely to drop. Gavin (Program Staff) at HealthNGO described that his main interest was in the accuracy of this prediction:

> "See when we are saying 85 percent [prediction accuracy], we are looking at that success rate and internal [call success rate] is a different success rate than what I'm saying. *I am looking at how many women are prevented from falling out of the program.* That would be my success rate…If she [the beneficiary] is not interested in listening

> to the call, or if she's not listening to the calls because she does not have network in her house, then it AI is not going to help."—Gavin (Program Staff)

Gavin was interested in the prediction accuracy, but an additional metric at the program level was the "call success rate" or how many of the calls placed through the AI program were answered by beneficiaries. The challenge with focusing on this metric, as Gavin conveyed, was that it depended on beneficiary behavior that seemed to be out of scope for what AI could enable. However, we saw how certain decisions, such as calling at specific times of the day, could increase the rate of calls being answered. A third success metric associated with the program that was of interest to the callers was the outcome of the calls that were picked up. Our focus group with TechOrg highlighted that not all of the above desired outcomes were necessarily tied to increased engagement. For instance, changing the timing of the mCare calls had not resulted in greater engagement from the beneficiaries, which may have been due to the uncertainties around schedules and the complexity of calling behavior and movement patterns that we detailed earlier.

5.4.2 Streamlining Data Flows. The introduction of AI had brought HealthNGO to streamline data flows. For instance, Pramila (Program Staff) stated, *"Many changes have come as a result of introducing some of the changes [in the AI program]. We introduced the ARN program as a result."* The ARN (Acquirer Reference Number) was a unique ID generated during the registration process and helped verify that a beneficiary's number was operational. It was introduced because of the challenges faced by hospital supervisors, as Leena shared, *"because in the TechOrg program, the calls would not go through"*. She further explained that this was because the phone numbers provided during registration might not have been correct or operational. Through the ARN program, beneficiaries were now required to place a missed call to opt into the program. They then received a text message with the ARN, which was recorded by HealthNGO.

5.4.3 Evaluating Impact. Due to the iterative nature of the intervention, both in terms of the ML model and associated workflows, program officers at HealthNGO had shared that they found it challenging to determine the impact that it was having in this context. One effect that they did note, however, was significant increase in overall engagement across the mCare program since the ML system had been introduced. This may have been a result of learnings and workflows incorporated in the course of implementing the program, such as incorporating feedback from beneficiaries, the ARN data flow, and offline follow-up interactions in hospitals, which had strengthened the mCare program overall.

Our focus group discussions with TechOrg conveyed that they were conducting a randomized controlled experiment, as a follow-up to the study of the previous ML model in use that showed a 30% increase in engagement compared to those who did not receive the intervention. TechOrg has been testing the effectiveness of two updated ML models towards increasing engagement in comparison to a round-robin approach (calling based on a systematic sequential basis). The initial stages of implementation found a relatively less significant difference between the performance of the round-robin and ML models, compared to the performance in the first impact evaluation. In our focus group discussions, we found that this had prompted stakeholders at HealthNGO and TechOrg to consider what may have been shaping model effectiveness. One change that they were already incorporating through dialogue with HealthNGO was taking call outcomes into consideration in the ML model.

6 DISCUSSION

Our analysis detailed the configuration, implementation, and evaluation of an AI intervention in public health, and how they shaped and were shaped by the interactions across diverse stakeholders.

We now present takeaways from our research, drawing attention to (1) the work done by the AI, (2) the work of building the human-AI partnership, and (3) the work of aligning program goals for the design and implementation of the AI system.

6.1 AI as Actor

Our analysis outlined how the integration of AI impacted program goals, design, and human interactions in the context of a public health intervention. Seeing AI as an *actor* with agency allows us to interrogate the work that it was doing in this context. Here we draw a connection to Actor-Network Theory (ANT), an analytical and methodological approach that views humans and nonhumans as having agency and playing an *equal* role in acting or participating in a network of relationships [53, 89]. There were several interconnected networks that AI acted on in our research, including (but not limited to) the organizational structures at HealthNGO and TechOrg, the relationships across callers and beneficiaries (which varied in strength across CCEs and hospital supervisors), and to a lesser extent, the relationships between beneficiaries and their families. Here we draw attention to some ways in which AI acted in this network of networks.

AI was initially introduced in the mCare program with the goal of increasing beneficiary engagement, to improve health outcomes overall. It sought to *optimize* limited resources for program staff at HealthNGO, by identifying beneficiaries who were disengaging from the program and could be targeted with a human intervention. This is a common approach in AI research, to support resource allocation in resource-constrained settings [22, 94]. With time, the AI model's targeting of beneficiaries became more complex as the understanding of beneficiary behavior became increasingly nuanced, facilitated by the experience of developing and deploying the models. The focus shifted to identifying beneficiaries who could benefit from a human intervention in diverse ways, such as by providing counseling and stopping calls to women who may have experienced a miscarriage or helping women navigate gendered access to the phone. AI thus took on the role of *advocate* by drawing the attention of callers and program staff to the challenges that a beneficiary might be experiencing. AI also became a *mediator,* enabling dialogue on program and design goals across stakeholders. This included discussions on the impact or success of the intervention at various levels—health outcomes for beneficiaries, increased program engagement for HealthNGO, and call outcomes for callers. We also found that the integration of AI stimulated conversations around existing workflows, and how they could be better aligned with the desired impact. These also had to be weighed against other priorities for HealthNGO, such as fairness, privacy, sustainability, caller motivation, among others.

AI systems might play similar roles in extended partnerships as part of public sector programs in other domains—as an optimizer of resources, advocate for target populations, mediator of dialogue across actors, and others. In particular, future work could further investigate the unique roles that AI could play as advocate and moderator, and how we might explicitly design for AI to enable dialogue across stakeholders. The networks within which an AI system is embedded could include institutions, policies, and other networked human and nonhuman actors that are crucial stakeholders in public sector programs. A focus on AI as an embedded actor with agency and the roles it plays can help identify where and how best AI might be integrated within broader ecosystems, and how human-AI partnerships can be configured to align with program goals, discussed further in the next section.

Rediet et al. have previously proposed that computing can offer four roles to address social change by serving as a diagnostic, formalizer, rebuttal, and synecdoche, which align closely with our conceptualization of AI as an advocate and mediator [1]. In other work, Suh et al. have described how AI can play a role as a social glue, by supporting co-creation among humans, and we see a similar effect enabled through the process of AI development itself [82]. However, our research

also draws attention to the *power* that AI held in our study context, by shaping the attention of and conversations among stakeholders. On one hand, this effect was being leveraged towards addressing equity. In particular, AI could play a role in drawing the attention of human actors in certain time-critical scenarios, such as in the case of miscarriage in our study context or for pregnancy and newborn complications. In such cases, the sense of urgency should also be communicated to the human actors intervening. On the other hand, we also raise concern around the ethics of leveraging AI for such roles, when model performance may be uncertain, program goals may not be straightforward, and AI could have an undue influence on shaping decisions, as several studies have pointed to the risk of [11, 43].

6.2 Configuring Human-AI Partnerships

We have thus far repeatedly discussed AI as a single entity in the context we studied. In reality, what was perceived to be an "AI intervention" was a set of human and nonhuman/technological actors working together. This included the ML model in the background developed by TechOrg, the web application used by the research staff at HealthNGO to distribute calls, the mobile application used by callers, as well as the calls made to the phones of beneficiaries, access to which had to be negotiated with family members. Each of these human and nonhuman actors was embedded in the ecosystem. A rich body of literature has emerged on human-AI collaboration in the HCI and FAccT (Fairness, Accountability, and Transparency) communities. Several studies have examined the role of humans-in-the-loop [78] or machine-in-the-loop [15, 28], among other conceptualizations of humans and AI systems working together. Mackeprang and Müller-Birn et al. have previously reflected on human-computer configuration design through their investigation of an intelligence system for collaborative ideation [56]. Their work conceptualizes differing levels of automation as a continuum, with the appropriate level to be determined based on the context [56]. We build on this perspective, drawing attention to the configuration that human-AI/ML partnership might entail as part of an ecosystem, and considering the effect of the human and the technology in varying degrees within the partnership, rather than trying to determine what is "in the loop."

In the current configuration of the human-AI partnership we studied, AI was largely situated in the background of a human intervention. AI helped optimize and direct limited resources and shaped the understanding of program and development staff at HealthNGO and TechOrg about beneficiary behavior. Its impact on the interactions between callers and beneficiaries was deliberately contained. This was a choice made to avoid influencing these interactions and preserving the autonomy and decision-making authority of callers. Our findings, however, indicated that the "hustling" the callers engaged in to connect with beneficiaries was complex and challenging, and differed across callers based on their role, experience, and expertise. To further strengthen AI's efficacy as advocate, the setup could be reconfigured for AI to have more influence on callers' decision-making, thereby supporting and easing their workflows. This could be done, for example, by making some of the beneficiary behavior visible to callers on the mobile application. We caution that this increased visibility would have to be managed carefully in technology design. Prior HCI research has highlighted how users can be susceptible to AI's suggestions and may defer to AI authority [27, 29, 43, 100]. A perspective on configuration also allows us to creatively combine human and technological capacities.

We also draw attention to the work that (re)configuration might entail. For instance, the current configuration of the human-AI partnership relied significantly on the assistance of the program and development staff at HealthNGO and TechOrg. It involved sharing anonymized caller data, generating the list of beneficiaries to call using the ML model, and then distributing the list across callers. Callers were also expected to download their list of assigned beneficiaries and upload call outcomes. This configuration was arrived at through dialogue across stakeholders to ensure

minimal work for the callers, and to support the transfer of the ML model to HealthNGO in the near future. Depending on the long-term goals of the intervention, the human assistance offered to the AI to make effective use of it may also have to be carefully configured. We next reflect on the role of dialogue in shaping the AI intervention.

6.3 Aligning Program Goals Across Stakeholders

Our findings detailed the various decisions involved when integrating AI in our study context, such as who to target with the intervention and when. In making these decisions, stakeholders had to engage in continual dialogue around program goals. Below we discuss two goals, as examples, that stood out in our research—fairness and intervention success. We consider how stakeholders tried to find alignment around these goals.

We found a shared, though not well-defined, understanding around *fairness* across stakeholders in this context. A fair ML model was perceived to be one that targeted populations that were marginal along one or more socioeconomic factors (e.g., income, education, caste, religion, migrant status). In practice, model fairness was constrained by the availability of reliable data for these dimensions. The populations targeted in an effort to be fair were also less likely to be able to listen to calls due to shared phone ownership, network issues, and more. This could affect caller motivation which was important for program sustainability, and might not have been fair to them. Fairness could also mean identifying those who were not a target of the intervention. For instance, identifying miscarriages was seen as important from an ethical standpoint, even if it reduced the number of enrolled participants in the program. Low call success rates could also come into conflict with what the program overall was optimizing for, which was beneficiary engagement. The various perspectives regarding who to be fair to and how, along with practical considerations around data availability and program sustainability, had to be balanced during ML development. Also, though fairness was discussed as a goal, what a fair model would look like was not explicitly defined. Instead, each iteration of the ML model was targeted towards making it *fairer*, through input from HealthNGO. Given that the end goal was to hand over the model to HealthNGO, one key challenge was determining when the model was "fair enough." Formalizing fairness has been a significant focus of the FAccT community [45, 83, 90], but this approach and limited emphasis on justice has been critiqued [26]. Our research points to the inherent tensions with targeting equity when working with multiple intersections of identity. Beyond formalizing fairness or justice as issues that can be fixed with a better model, there is a need for AI practitioners to grapple with the messy, shifting, and embedded nature of inequities. The model itself may play a limited contributing role as part of a broader effort (such a the mCare intervention overall) to improve equity.

We also found that stakeholders approached *intervention success* in multiple ways in this setting. Success could mean high predictive accuracy of the model before deployment. It could also mean increased engagement, increase in success rates, or optimization of certain call outcomes after deployment. Given the time-sensitive nature of certain health concerns (such as pregnancy complications or miscarriage), the timeliness of delivering an intervention could be another aspect of performance to consider. The multiple approaches point to a subtle difference between model efficacy/outcomes and overall program efficacy/outcomes. While typical metrics around predictive accuracy or shift in engagement could help understand model performance, secondary outcomes such as call success rates and call outcomes could help understand if and how the system was reaching targeted populations. For instance, a focus on measuring model efficacy through increase in engagement does not capture the nature of engagement, which call outcomes would reflect. In this context, a modest increase in engagement from beneficiaries from marginalized backgrounds might be more desirable than a higher increase in engagement of those who would benefit less from the program. This may have been reflected in the latest ML model which was comparable in terms of

increasing engagement to a round-robin approach, but may have been selecting more beneficiaries who were more marginalized and less engaged to begin with. We also saw that decisions around the program design, such as when to call beneficiaries, could shape model performance.

Yang et al. have described why AI is uniquely difficult to design for because of the uncertainty surrounding its capabilities and output complexity [99]. We found this in our context as well, and an iterative and collaborative process was critical in working with such uncertainties. For AI researchers and developers increasingly working with the public sector, one approach to resolve uncertainties could be to engage in Design-Based Implementation Research (DBIR), a methodology that emerged from practice-oriented research in the education domain and emphasizes an iterative and collaborative approach [23]. It operates on four core principles: "(1) a focus on persistent problems of practice from multiple stakeholders' perspectives; (2) a commitment to iterative, collaborative design; (3) a concern with developing theory and knowledge related to both learning and implementation through systematic inquiry; and (4) a concern with developing capacity for sustaining change in systems." [23, 51]. Prior work has discussed how DBIR can unite HCI researchers and practitioners in their shared commitments "towards informed practice", while allowing them to honor their primary commitments [51]. DBIR is one example of an approach that could help align stakeholders around program goals when integrating AI into real-world settings.

7 LIMITATIONS AND FUTURE WORK

For an AI system such as the one our paper examines, there are many and diverse stakeholders and stakeholder perspectives. We investigated AI integration from the perspectives of those implementing and using the AI system, including callers (CCEs and hospital supervisors), program staff at HealthNGO, and the development team at TechOrg. We chose not to directly engage with beneficiaries because of early interactions with HealthNGO that revealed that on account of their decision to explicitly integrate AI in the background of their public health program, beneficiaries only interfaced with human actors over a phone call and had no interactions with the AI system. The phone calls to beneficiaries based on the predictions of the AI model were also similar to the existing *37-week* and *missed call* initiatives. These were conscious design choices to ensure that an individual's experience in the mCare program was no different after the introduction of AI. Though our findings describe implementation and design considerations around an AI system developed to support a specific public health program in India, our work has broader implications as similar systems for resource allocation are introduced in other resource-constrained settings. We hope our research can inform the ethnographic study of AI systems in other contexts, and that our discussion on the role and configuration of AI systems can shape how they are implemented in health settings and public sector programs. Future research could build on our findings, depending on the beneficiaries and intervention contexts targeted, and potentially benefit from adapting methods towards eliciting community perspectives on AI interventions.

8 CONCLUSION

AI systems are increasingly being integrated into public sector programs for decision-support and distribution of limited resources. Human-Computer Interaction (HCI) research has begun to examine how such systems may be designed appropriately when targeting underserved populations. Our research offers an ethnographic study of a large-scale real-world deployment of an AI system for resource allocation in a call-based maternal and child health information delivery program in India. In this paper, we began by presenting the *what* or the program definition of the AI intervention, before uncovering complexities around determining *who* benefits, *how* the human-AI collaboration is managed, *when* intervention must take place in alignment with various priorities, and *why* the AI is sought, for what purpose. Our paper draws attention to the work done by the AI

(as actor), the work of building the human-AI partnership (with multiple, diverse stakeholders), and the work of aligning program goals for design and implementation (through continual dialogue across stakeholders).

ACKNOWLEDGMENTS

We are grateful to our anonymous reviewers for their encouragement and feedback, and to our study participants and partner HealthNGO for sharing their valuable time and expertise. This material is based upon work supported by the National Science Foundation under Grant No. 2047726. We also thank Google.org and Google University Relations for their funding towards this project. We would like to thank Milind Tambe, Aparna Taneja, Shresth Verma, James Wexler, and the Tandem Lab at Georgia Tech for their generous feedback on the paper.

REFERENCES

[1] Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G Robinson. 2020. Roles for computing in social change. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 252–260.

[2] Syed Ishtiaque Ahmed, Md Romael Haque, Jay Chen, and Nicola Dell. 2017. Digital Privacy Challenges with Shared Mobile Phone Use in Bangladesh. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 17.

[3] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, and others. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 3.

[4] Asbjørn Ammitzbøll Flügge, Thomas Hildebrandt, and Naja Holten Møller. 2021. Street-Level Algorithms and AI in Bureaucratic Decision-Making: A Caseworker Perspective. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 1–23. https://doi.org/10.1145/3449114

[5] Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S Weld. 2021. Is the most accurate ai the best teammate? optimizing ai for teamwork. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 11405–11414.

[6] 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.

[7] Rama Baru, Arnab Acharya, Sanghmitra Acharya, AK Shiva Kumar, and K Nagaraj. 2010. Inequities in access to health services in India: caste, class and region. *Economic and political Weekly* (2010), 49–58.

[8] 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*. ACM, Honolulu HI USA, 1–12. https://doi.org/10.1145/3313831.3376718

[9] Anna Brown, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. 2019. Toward Algorithmic Accountability in Public Services: A Qualitative Study of Affected Community Perspectives on Algorithmic Decision-making in Child Welfare Services. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, 41:1–41:12. https://doi.org/10.1145/3290605.3300271 event-place: Glasgow, Scotland Uk.

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

[11] Shiye Cao and Chien-Ming Huang. 2022. Understanding User Reliance on AI in Assisted Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 471 (nov 2022), 23 pages. https://doi.org/10.1145/3555572

[12] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. 2017. Revolt: Collaborative Crowdsourcing for Labeling Machine Learning Datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, New York, NY, USA, 2334–2346. https://doi.org/10.1145/3025453.3026044

[13] Kathy Charmaz. 2006. *Constructing grounded theory: A practical guide through qualitative analysis*. sage.

[14] Janet X Chen, Allison McDonald, Yixin Zou, Emily Tseng, Kevin A Roundy, Acar Tamersoy, Florian Schaub, Thomas Ristenpart, and Nicola Dell. 2022. Trauma-Informed Computing: Towards Safer Technology Experiences for All. In *CHI Conference on Human Factors in Computing Systems*. 1–20.

[15] Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A Smith. 2018. Creative writing with a machine in the loop: Case studies on slogans and stories. In *23rd International Conference on Intelligent User Interfaces*.

329–340.
[16] Anupam Datta, Shayak Sen, and Yair Zick. 2016. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE symposium on security and privacy (SP)*. IEEE, 598–617.
[17] 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*. ACM, 2220–2232.
[18] Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. 2017. UX Design Innovation: Challenges for Working with Machine Learning as a Design Material. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, New York, NY, USA, 278–288. https://doi.org/10.1145/3025453.3025739
[19] 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.
[20] Upol Ehsan, Ranjit Singh, Jacob Metcalf, and Mark Riedl. 2022. The Algorithmic Imprint. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 1305–1317.
[21] Laura Elizabeth Ellington, Irene Najjingo, Margaret Rosenfeld, James W Stout, Stephanie A Farquhar, Aditya Vashistha, Bridget Nekesa, Zaituni Namiya, Agatha J Kruse, Richard Anderson, et al. 2021. Health workers' perspectives of a mobile health tool to improve diagnosis and management of paediatric acute respiratory illnesses in Uganda: a qualitative study. *BMJ open* 11, 7 (2021), e049708.
[22] Fei Fang, Thanh Hong Nguyen, Rob Pickles, Wai Y Lam, Gopalasamy R Clements, Bo An, Amandeep Singh, Milind Tambe, Andrew Lemieux, et al. 2016. Deploying PAWS: Field Optimization of the Protection Assistant for Wildlife Security.
[23] Barry J Fishman, William R Penuel, Anna-Ruth Allen, Britte Haugan Cheng, and NORA Sabelli. 2013. Design-based implementation research: An emerging model for transforming the relationship of research and practice. *Teachers College Record* 115, 14 (2013), 136–156.
[24] Jacquelline Fuller. 2019. Here are the grantees of the Google AI Impact Challenge. https://www.blog.google/outreach-initiatives/google-org/ai-impact-challenge-grantees/
[25] Ya'akov Gal, Avi Segal, Ece Kamar, Eric Horvitz, Chris Lintott, and Mike Walmsley. 2022. A new Workflow for Human-AI Collaboration in Citizen Science. In *Proceedings of the 2022 ACM Conference on Information Technology for Social Good*. 89–95.
[26] Ben Gansky and Sean McDonald. 2022. CounterFAccTual: How FAccT undermines its organizing principles. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 1982–1992.
[27] Susanne Gaube, Harini Suresh, Martina Raue, Alexander Merritt, Seth J. Berkowitz, Eva Lermer, Joseph F. Coughlin, John V. Guttag, Errol Colak, and Marzyeh Ghassemi. 2021. Do as AI say: susceptibility in deployment of clinical decision-aids. *npj Digital Medicine* 4, 1 (Feb. 2021), 1–8. https://doi.org/10.1038/s41746-021-00385-9
[28] Ben Green and Yiling Chen. 2019. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the conference on fairness, accountability, and transparency*. 90–99.
[29] Ben Green and Yiling Chen. 2021. Algorithmic risk assessments can alter human decision-making processes in high-stakes government contexts. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–33.
[30] Jonathan Grudin and Richard Jacques. 2019. Chatbots, humbots, and the quest for artificial general intelligence. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–11.
[31] Mukesh Hamal, Marjolein Dieleman, Vincent De Brouwere, and Tjard de Cock Buning. 2020. Social determinants of maternal health: a scoping review of factors influencing maternal mortality and maternal health service use in India. *Public Health Reviews* 41, 1 (2020), 1–24.
[32] 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.
[33] Fred Hohman, Kanit Wongsuphasawat, Mary Beth Kery, and Kayur Patel. 2020. Understanding and Visualizing Data Iteration in Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.
[34] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, III, Miro Dudik, and Hanna Wallach. 2019. Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, 600:1–600:16. https://doi.org/10.1145/3290605.3300830 event-place: Glasgow, Scotland Uk.
[35] Naja Holten Møller, Irina Shklovski, and Thomas T. Hildebrandt. 2020. Shifting Concepts of Value: Designing Algorithmic Decision-Support Systems for Public Services. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*. ACM, Tallinn Estonia, 1–12. https://doi.org/10.1145/3419249.3420149

[36] Azra Ismail and Neha Kumar. 2019. Empowerment on the Margins: The Online Experiences of Community Health Workers. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 99.
[37] 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.
[38] Maia Jacobs, Jeffrey He, Melanie F. Pradier, Barbara Lam, Andrew C. Ahn, Thomas H. McCoy, Roy H. Perlis, Finale Doshi-Velez, and Krzysztof Z. Gajos. 2021. Designing AI for Trust and Collaboration in Time-Constrained Medical Decisions: A Sociotechnical Lens. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3411764.3445385
[39] Maurice Jakesch, Zana Buçinca, Saleema Amershi, and Alexandra Olteanu. 2022. How Different Groups Prioritize Ethical Values for Responsible AI. *arXiv preprint arXiv:2205.07722* (2022).
[40] Maurice Jakesch, Megan French, Xiao Ma, Jeffrey T Hancock, and Mor Naaman. 2019. AI-mediated communication: How the perception that profile text was written by AI affects trustworthiness. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.
[41] Yonggeol Jo, Minwoo Kim, and Kyungsik Han. 2019. How Do Humans Assess the Credibility on Web Blogs: Qualifying and Verifying Human Factors with Machine Learning. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.
[42] Young Hun Jung and Ambuj Tewari. 2019. Regret bounds for thompson sampling in episodic restless bandit problems. *Advances in Neural Information Processing Systems* 32 (2019).
[43] Shivani Kapania, Oliver Siy, Gabe Clapper, Azhagu Meena SP, and Nithya Sambasivan. 2022. "Because AI is 100% right and safe": User Attitudes and Sources of AI Authority in India. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3491102.3517533
[44] Naveena Karusala, Apoorva Bhalla, and Neha Kumar. 2019. Privacy, Patriarchy, and Participation on Social Media. In *Proceedings of the 2019 on Designing Interactive Systems Conference* (San Diego, CA, USA) *(DIS '19)*. ACM, ACM, New York, NY, USA, 511–526. https://doi.org/10.1145/3322276.3322355 event-place: San Diego, CA, USA.
[45] Maximilian Kasy and Rediet Abebe. 2021. Fairness, equality, and power in algorithmic decision-making. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 576–586.
[46] Anna Kawakami, Venkatesh Sivaraman, Hao-Fei Cheng, Logan Stapleton, Yanghuidi Cheng, Diana Qing, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. 2022. Improving Human-AI Partnerships in Child Welfare: Understanding Worker Practices, Challenges, and Desires for Algorithmic Decision Support. In *CHI Conference on Human Factors in Computing Systems*. 1–18.
[47] Anna Kawakami, Venkatesh Sivaraman, Logan Stapleton, Hao-Fei Cheng, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. 2022. "Why Do I Care What's Similar?" Probing Challenges in AI-Assisted Child Welfare Decision-Making through Worker-AI Interface Design Concepts. In *Designing Interactive Systems Conference*. ACM, Virtual Event Australia, 454–470. https://doi.org/10.1145/3532106.3533556
[48] Pranav Khadpe, Ranjay Krishna, Li Fei-Fei, Jeffrey T Hancock, and Michael S Bernstein. 2020. Conceptual metaphors impact perceptions of human-ai collaboration. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–26.
[49] Jackson A Killian, Bryan Wilder, Amit Sharma, Vinod Choudhary, Bistra Dilkina, and Milind Tambe. 2019. Learning to prescribe interventions for tuberculosis patients using digital adherence data. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2430–2438.
[50] Neha Kumar. 2015. The gender-technology divide or perceptions of non-use? *First Monday* 20, 11 (2015).
[51] Neha Kumar and Nicola Dell. 2018. Towards informed practice in HCI for development. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–20.
[52] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a science of human-ai decision making: a survey of empirical studies. *arXiv preprint arXiv:2112.11471* (2021).
[53] Bruno Latour. 2007. *Reassembling the social: An introduction to actor-network-theory*. Oup Oxford.
[54] Min Hun Lee, Daniel P. Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 2021. A Human-AI Collaborative Approach for Clinical Decision Making on Rehabilitation Assessment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3411764.3445472
[55] Min Kyung Lee and Katherine Rich. 2021. Who Is Included in Human Perceptions of AI?: Trust and Perceived Fairness around Healthcare AI and Cultural Mistrust. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–14. https://doi.org/10.1145/3411764.3445570
[56] Maximilian Mackeprang, Claudia Müller-Birn, and Maximilian Timo Stauss. 2019. Discovering the sweet spot of human-computer configurations: A case study in information extraction. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–30.

[57] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[58] George E Marcus. 1995. Ethnography in/of the world system: The emergence of multi-sited ethnography. *Annual review of anthropology* (1995), 95–117.

[59] Aditya Mate, Lovish Madaan, Aparna Taneja, Neha Madhiwalla, Shresth Verma, Gargi Singh, Aparna Hegde, Pradeep Varakantham, and Milind Tambe. 2022. Field Study in Deploying Restless Multi-Armed Bandits: Assisting Non-profits in Improving Maternal and Child Health. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 11 (June 2022), 12017–12025. https://doi.org/10.1609/aaai.v36i11.21460

[60] Vineet Nair, Kritika Prakash, Michael Wilbur, Aparna Taneja, Corrine Namblard, Oyindamola Adeyemo, Abhishek Dubey, Abiodun Adereni, Milind Tambe, and Ayan Mukhopadhyay. 2022. ADVISER: AI-Driven Vaccination Intervention Optimiser for Increasing Vaccine Uptake in Nigeria. *arXiv preprint arXiv:2204.13663* (2022).

[61] Siddharth Nishtala, Harshavardhan Kamarthi, Divy Thakkar, Dhyanesh Narayanan, Anirudh Grama, Ramesh Padmanabhan, Neha Madhiwalla, Suresh Chaudhary, Balaraman Ravindra, and Milind Tambe. 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).

[62] Changhoon Oh, Jungwoo Song, Jinhan Choi, Seonghyeon Kim, Sungwoo Lee, and Bongwon Suh. 2018. I lead, you help but only with enough details: Understanding user experience of co-creation with artificial intelligence. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

[63] 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)*. ACM, Seattle WA USA, 439–452. https://doi.org/10.1145/3530190.3534802

[64] 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 (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–20. https://doi.org/10.1145/3411764.3445420

[65] Cecilia Panigutti, Andrea Beretta, Fosca Giannotti, and Dino Pedreschi. 2022. Understanding the impact of explanations on advice-taking: a user study for AI-based clinical Decision Support Systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3491102.3502104

[66] Andrea Papenmeier, Dagmar Kern, Daniel Hienert, Yvonne Kammerer, and Christin Seifert. 2022. How Accurate Does It Feel?–Human Perception of Different Types of Classification Mistakes. In *CHI Conference on Human Factors in Computing Systems*. 1–13.

[67] Samir Passi and Steven J Jackson. 2018. Trust in Data Science: Collaboration, Translation, and Accountability in Corporate Data Science Projects. 2 (2018), 136: 1–136: 28. Issue CSCW.

[68] Pintu Paul and Pradip Chouhan. 2020. Socio-demographic factors influencing utilization of maternal health care services in India. *Clinical Epidemiology and Global Health* 8, 3 (2020), 666–670.

[69] Marissa Radensky, Doug Downey, Kyle Lo, Zoran Popovic, and Daniel S Weld. 2022. Exploring the Role of Local and Global Explanations in Recommender Systems. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–7.

[70] Charvi Rastogi, Yunfeng Zhang, Dennis Wei, Kush R Varshney, Amit Dhurandhar, and Richard Tomsett. 2020. Deciding fast and slow: The role of cognitive biases in ai-assisted decision-making. *arXiv preprint arXiv:2010.07938* (2020).

[71] Negar Rostamzadeh, Diana Mincu, Subhrajit Roy, Andrew Smart, Lauren Wilcox, Mahima Pushkarna, Jessica Schrouff, Razvan Amironesei, Nyalleng Moorosi, and Katherine Heller. 2022. Healthsheet: Development of a Transparency Artifact for Health Datasets. *arXiv preprint arXiv:2202.13028* (2022).

[72] 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 (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 315–328. https://doi.org/10.1145/3442188.3445896

[73] 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*. 315–328.

[74] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*.

[75] Devansh Saxena, Karla Badillo-Urquiola, Pamela J Wisniewski, and Shion Guha. 2020. A human-centered review of algorithms used within the US child welfare system. In *Proceedings of the 2020 CHI Conference on Human Factors in*

Computing Systems. 1–15.
[76] Devansh Saxena, Seh Young Moon, Dahlia Shehata, and Shion Guha. 2022. Unpacking Invisible Work Practices, Constraints, and Latent Power Relationships in Child Welfare through Casenote Analysis. In *CHI Conference on Human Factors in Computing Systems*. 1–22.
[77] Morgan Klaus Scheuerman, Emily Denton, and Alex Hanna. 2021. Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development. *arXiv preprint arXiv:2108.04308* (2021).
[78] Ben Shneiderman. 2020. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction* 36, 6 (2020), 495–504.
[79] C Estelle Smith, Bowen Yu, Anjali Srivastava, Aaron Halfaker, Loren Terveen, and Haiyi Zhu. 2020. Keeping Community in the Loop: Understanding Wikipedia Stakeholder Values for Machine Learning-Based Systems. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.
[80] Thomas N Smyth, Satish Kumar, Indrani Medhi, and Kentaro Toyama. 2010. Where there's a will there's a way: mobile media sharing in urban india. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM, 753–762.
[81] Minhyang Suh, Emily Youngblom, Michael Terry, and Carrie J Cai. 2021. AI as social glue: uncovering the roles of deep generative AI during social music composition. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–11.
[82] Minhyang (Mia) Suh, Emily Youngblom, Michael Terry, and Carrie J Cai. 2021. AI as Social Glue: Uncovering the Roles of Deep Generative AI during Social Music Composition. 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 582, 11 pages. https://doi.org/10.1145/3411764.3445219
[83] Harini Suresh and John Guttag. 2021. A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and access in algorithms, mechanisms, and optimization*. 1–9.
[84] Harini Suresh, Rajiv Movva, Amelia Lee Dogan, Rahul Bhargava, Isadora Cruxen, Angeles Martinez Cuba, Guilia Taurino, Wonyoung So, and Catherine D'Ignazio. 2022. Towards Intersectional Feminist and Participatory ML: A Case Study in Supporting Feminicide Counterdata Collection. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 667–678. https://doi.org/10.1145/3531146.3533132
[85] Harini Suresh, Rajiv Movva, Amelia Lee Dogan, Rahul Bhargava, Isadora Cruxên, Ángeles Martinez Cuba, Guilia Taurino, Wonyoung So, and Catherine D'Ignazio. 2022. Towards Intersectional Feminist and Participatory ML: A Case Study in Supporting Feminicide Counterdata Collection. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 667–678.
[86] Divy Thakkar, Azra Ismail, Pratyush Kumar, Alex Hanna, Nithya Sambasivan, and Neha Kumar. 2022. When is Machine Learning Data Good?: Valuing in Public Health Datafication. In *CHI Conference on Human Factors in Computing Systems*. 1–16.
[87] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3174014
[88] Sarah Theres Völkel, Renate Haeuslschmid, Anna Werner, Heinrich Hussmann, and Andreas Butz. 2020. How to Trick AI: Users' Strategies for Protecting Themselves from Automatic Personality Assessment. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.
[89] Geoff Walsham. 1997. Actor-network theory and IS research: current status and future prospects. *Information systems and qualitative research* (1997), 466–480.
[90] Angelina Wang, Vikram V Ramaswamy, and Olga Russakovsky. 2022. Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 336–349.
[91] Dakuo Wang, Liuping Wang, Zhan Zhang, Ding Wang, Haiyi Zhu, Yvonne Gao, Xiangmin Fan, and Feng Tian. 2021. "Brilliant AI Doctor" in Rural Clinics: Challenges in AI-Powered Clinical Decision Support System Deployment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3411764.3445432
[92] Dakuo Wang, Justin D Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. 2019. Human-AI collaboration in data science: Exploring data scientists' perceptions of automated AI. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.
[93] 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.

[94] Bryan Wilder, Marie Charpignon, Jackson A Killian, Han-Ching Ou, Aditya Mate, Shahin Jabbari, Andrew Perrault, Angel N Desai, Milind Tambe, and Maimuna S Majumder. 2020. Modeling between-population variation in COVID-19 dynamics in Hubei, Lombardy, and New York City. *Proceedings of the National Academy of Sciences* 117, 41 (2020), 25904–25910.

[95] Amulya Yadav, Bryan Wilder, Eric Rice, Robin Petering, Jaih Craddock, Amanda Yoshioka-Maxwell, Mary Hemler, Laura Onasch-Vera, Milind Tambe, and Darlene Woo. 2018. Bridging the gap between theory and practice in influence maximization: Raising awareness about HIV among homeless youth.. In *IJCAI*. 5399–5403.

[96] Deepika Yadav, Prerna Malik, Kirti Dabas, and Pushpendra Singh. 2019. Feedpal: Understanding Opportunities for Chatbots in Breastfeeding Education of Women in India. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 170.

[97] Jing Nathan Yan, Ziwei Gu, Hubert Lin, and Jeffrey M Rzeszotarski. 2020. Silva: Interactively Assessing Machine Learning Fairness Using Causality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[98] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. In *Proceedings of the 2020 chi conference on human factors in computing systems*. 1–13.

[99] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, 279:1–279:12. https://doi.org/10.1145/3290605.3300509 event-place: Glasgow, Scotland Uk.

[100] Yue You, Yubo Kou, Xianghua(Sharon) Ding, and Xinning Gui. 2021. The Medical Authority of AI: A Study of AI-enabled Consumer-Facing Health Technology. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3411764.3445657

[101] Rui Zhang, Nathan J McNeese, Guo Freeman, and Geoff Musick. 2021. " An Ideal Human" Expectations of AI Teammates in Human-AI Teaming. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–25.

Received January 2023; revised April 2023; accepted May 2023