Nurturing Capabilities: Unpacking the Gap in Human-Centered Evaluations of AI-Based Systems

Aman Khullar
School of Interactive Computing
Georgia Institute of Technology
Atlanta, Georgia, USA
akhullar8@gatech.edu

Nikhil Nalin
Noora Health
Bangalore, India
nikhil@noorahealth.org

Abhishek Prasad
Noora Health
Bangalore, India
abhishek.prasad@noorahealth.org

Ann John Mampilli
Noora Health
Bangalore, India
ann@noorahealth.org

Neha Kumar
Georgia Tech
Atlanta, Georgia, USA
neha.kumar@cc.gatech.edu

Abstract

Human-Computer Interaction (HCI) scholarship has studied how Artificial Intelligence (AI) can be leveraged to support care work(ers) by recognizing, reducing, and redistributing workload. Assessment of AI's impact on workers requires scrutiny and is a growing area of inquiry within human-centered evaluations of AI. We add to these conversations by unpacking the sociotechnical gap between the broader aspirations of workers from an AI-based system and the narrower existing definitions of success. We conducted a mixed-methods study and drew on Amartya Sen's Capability Approach to analyze the gap. We shed light on the social factors—on top of performance on evaluation metrics—that guided the AI model choice and determined whose wellbeing must be evaluated while conducting such evaluations. We argue for assessing broader achievements enabled through AI's use when conducting human-centered evaluations of AI. We discuss and recommend the dimensions to consider while conducting such evaluations.

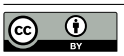
CCS Concepts

• **Human-centered computing → Empirical studies in HCI**; **HCI design and evaluation methods**.

Keywords

Social Determinants of AI Model Choice; Human Infrastructure; Aspirations; Mixed-Methods; DBIR; Capability Approach

ACM Reference Format:
Aman Khullar, Nikhil Nalin, Abhishek Prasad, Ann John Mampilli, and Neha Kumar. 2025. Nurturing Capabilities: Unpacking the Gap in Human-Centered Evaluations of AI-Based Systems. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3706598.3713278



CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713278

1 INTRODUCTION

The demanding nature of care work[1] is documented across fields like economics [32], gender studies [85], and Human-Computer Interaction (HCI) [46]. Shahra Razavi argues, "Good quality care, whether paid or unpaid, is very labour intensive" [85]. Feminist scholars across these fields call for automation to support reducing burdensome tasks in care work [32] while balancing the scaling of care services with the wellbeing of care workers [47]. In recent times, HCI scholars have examined Artificial Intelligence's (AI) potential to develop this automation and support the care workers [40, 73]. Human-centered evaluation of the impact of AI on the (care) workers is a relatively new area of study in HCI [114], and amidst these growing conversations is where we situate our work.

Human-centered evaluations of AI have assessed the benefits and challenges of AI-based system deployments and shed light on the gap between user expectations and field deployment of such systems [11]. HCI scholars are revisiting methodological approaches to conduct such evaluations in light of technical advancements in Large Language Models (LLMs) in AI [62, 113]. We build on conversations seeking to incorporate human-centered perspectives when asking "*what* to evaluate and *how* to evaluate & audit LLMs" [113]. We add to the calls arguing to conceptualize AI model evaluation as "narrowing the socio-technical gap" [62] based on the long-standing understanding of the sociotechnical gap between the social requirements of technology users and the technical feasibility of the technology used [2]. We add nuance to what determines the *technical feasibility* of AI models like LLMs and what should be considered when examining the *social requirements* of users of AI-based systems. In our work, we ask the research question, *"What determines the AI model choice, and what should HCI researchers and practitioners evaluate when conducting human-centered evaluations of AI-based systems?"*

We examine this question in the context of high-stakes domains[2] like health, education, and social justice. To conduct our inquiry, we collaborated with Noora Health India Private Limited, a wholly

[1]The International Labor Organization includes all the activities in education, health, social work, and unpaid domestic work as part of care work [76].
[2]We define high-stakes domain based on the definition offered by Sambasivan et al.—the domains that involve safety, wellbeing, and stakes (like road safety, credit assessment) of individuals [90].

owned subsidiary of Noora Health US (a 501(c)(3) non-profit organization) and their implementing partner, the YosAid Innovation Foundation[3]. We refer to Noora Health India Private Limited and their implementing partner as *HealthNGO* hereafter. The HealthNGO operates a mobile-based service to answer health-related queries from community members seeking care (*care recipients* hereafter) in Bangladesh, India, and Indonesia. The recipients' questions are answered by licensed health information workers employed by the HealthNGO. The HealthNGO sought automation to support its health information workers, i.e., their frontline care workers, who were experiencing a high workload due to the expanding service scale. We collaborated with HealthNGO's India-based operations to design an AI-based intervention to manage the workload of its care workers. We engaged with the Design-Based Implementation Research (DBIR) methodology [28] to create a design probe [30], developed iteratively through a multi-phase deployment. We conducted a mixed-method inquiry to assess its impact on the workers. In our collaboration with the HealthNGO, spanning nine months, we collected data through observations, interviews, our probe's usage logs, chat logs of care workers with recipients, and a focus group discussion. We drew on Sen's capability approach [94]—a theoretical framework to measure an individual's wellbeing—to analyze our data and assess our intervention's role in supporting care workers to achieve what they *needed* (to be successful at work) and what they *wanted* (as part of their broader aspirations at work) from the intervention.

We found a gap between what the care workers aspired to achieve through technology and how our human-centered evaluation method defined success. We unpack the reason for this gap and shed light on the social factors that determined the technical feasibility of our intervention and the broader social requirements of the care workers from our intervention. On top of an AI model's performance on evaluation metrics, access, financial incentives, and human resources required for the implementation effort determined the specific AI model, i.e., LLMs, chosen for the intervention. Deploying the AI-based system in a high-stakes public health domain required our intervention to incorporate additional human actors to mitigate AI fallibility [42]. The addition of human actors in our intervention complicated whose wellbeing must be assessed when conducting human-centered evaluations. Our quantitative assessment of workload reduction on the workers and supporting them to answer more recipient queries, upskilling, and dividing work through AI fell short of workers' broader aspirations for which they wanted to use the technology. We argue that human-centered evaluations should focus on an AI's ability to expand human capabilities—the substantive freedom AI facilitates in helping individuals achieve their aspirations. Our argument contrasts with a relatively narrow focus on assessing AI's ability to accumulate and augment human capital. We end our paper by discussing the *sociotechnical*, *ecological*, and *individual* dimensions that should be evaluated when conducting human-centered evaluations with a broader focus on expanding human capabilities. We now describe related prior work in which we situate our study, followed by details on our research methods and findings from our study, and then discuss the implications of our study for the HCI community.

[3]Nikhil and Abhishek are part of Noora Health India Private Limited and Ann is part of YosAid Innovation Foundation.

2 RELATED WORK

We situate our work within three key areas in HCI. We describe how researchers and practitioners choose to leverage a specific AI model, conduct human-centered evaluations, and design a human infrastructure to support AI. We contribute to these three areas by drawing from implementation and analytical frameworks from the fields of education and global development, respectively.

2.1 AI System Evaluation and Model Choice

Model evaluation is fundamental to implementing an AI-based system [104]. The evaluation comprises assessing the performance of the AI model using some *performance metrics* (like accuracy, precision, recall, F1 score, and false negative rate) [72] on evaluation dataset(s), which may be self-curated or openly available *benchmark datasets* (like SQUAD [82] and ImageNet [88]). An AI model's performance on such datasets measured through performance metrics typically determines which model will be chosen while developing an AI-based system [8, 11, 51]. Recent progress in LLMs has sparked renewed interest among AI and HCI researchers in designing frameworks and methodologies that can help compare the suitability of different AI models for a specific application [9, 109]. These studies argue for choosing an AI model based on multi-metric evaluations that also assess the societal impact of this technology (like carbon emissions, disinformation, financial cost, and representational harms) along with model performance on application tasks [44, 61, 100].

We build on studies investigating the societal impacts of AI-based systems. We align most closely with Barocas et al., studying the choices made by researchers and practitioners when conducting evaluations of AI-based systems [10]. They argue that researchers and practitioners make several critical choices while evaluating the impact of AI-based systems on people with different identities (like race and gender) and that it is crucial to understand those choices. These choices include asking what the goal of the evaluation should be, what factors should be focused on during the evaluation, when and where to conduct the evaluation, and who should conduct the evaluation and how. Dow et al. built on this work to incorporate the advancements in LLMs and proposed a set of dimensions that can help capture the critical choices while evaluating generative AI-based systems [25]. We expand on these studies and ask a more fundamental question of why researchers and practitioners make these choices. We study how *democratizing AI* by increasing access to AI technologies like LLMs [92] and reducing social barriers like financial cost and implementation effort [44] affects choices available to practitioners looking to leverage AI in their work. We broaden the understanding of what choices are substantively available to the AI model developers working with limited financial resources in a high-stakes domain.

2.2 Human-Centered Evaluations of AI

Moving beyond AI model evaluations, HCI researchers have argued for the need to conduct human-centered evaluations of AI-based

systems [11, 62, 114]. Beede et al. presented one of the first human-centered evaluations of an AI-based system and unpacked the negative impact of the system deployment environment on model performance and system usage in a clinical setting [11]. Building on this work, recent works have investigated the human perception and experiences of using such AI-based systems in chronic eye disease screening [78, 99] and if such systems are able to fulfill the emotional needs [66] and work-related needs [63] of individuals using the systems.

In a CHI 2024 workshop [114], researchers proposed to "rethink *what* to evaluate and *how* to evaluate & audit LLM." They built on Liao and Xiao's argument of viewing AI model evaluation as "narrowing the socio-technical gap" [62] and proposed a call to action for HCI and AI researchers and practitioners working on/with LLM-based systems to ask questions like "*Who* should be involved in evaluating and auditing LLMs? What are their needs and goals?" We build on this call and offer one answer to what should be evaluated when conducting human-centered evaluations of AI-based systems. We argue that HCI researchers and practitioners should expand the focus from needs and assess people's aspirations, i.e., what people *wanted* to achieve through leveraging an AI-based system instead of what people *needed* from the system usage. We contribute to studies investigating what it means to be human-centered in human-centered AI [17, 18]. We build on arguments focusing on human aspirations instead of needs in the design of human-centered AI [16]. We argue for bringing similar perspectives to the human-centered evaluations of AI-based systems by highlighting the gap between what our participants wanted to achieve through the use of the system and what they needed to achieve to be efficient and skilled in their work.

2.3 Human Infrastructure of AI in Public Health

The critical and often invisible work of human actors involved in making an AI system work is an active area of research [33, 86]. From training data preparation to output verification, humans play the role of *trainers, verifiers,* and *imitators* in supporting *micro-work* required for the real-world AI application to function [69, 107]. Passi and Sengers have unpacked how the different actors in this infrastructure negotiate to determine how, why, and what an AI-based system can achieve [79]. Elish and Watkins further argue that AI interventions should always be conceptualized as sociotechnical systems that require *repair work* for the intervention to be effective [27]. Due to the high impact of data quality on the output of the AI system [34, 68, 90], researchers have critically investigated the processing pipeline and the various actors involved in data production [31]. Prior studies have unpacked who performs *data work* [71, 103] and its high under-valuation [90, 91] across domains, including healthcare [15, 80, 81]. The different actors also play an active role in shaping the data [70] and its associated valuation during data transfers [106].

Verification workers—who verify the veracity of the output of an AI model [107]—are critical in a high-stakes domain like public health to mitigate AI's innate fallibility [7, 42, 59] from mitigating harms like non-evidentiary health-related information [7]. In line with recommendations from other domains (like seeking help from a human partner during a robot-induced error in human-robot interaction) [36], research in AI-based models in public health applications has highlighted the role of human *helpers* to verify the veracity of model output [83, 115]. The verification workers may be doctors [110] or healthcare workers [4]. We build on such studies leveraging verification workers' assets (like critical thinking skills) to identify and mitigate AI fallibility. Similar to prior work introducing and highlighting the tasks of verification workers—like the *overreaders* who tracked the recipients missed by the AI-based system's false predictions and required clinical attention [11]—we introduced verification workers to mitigate AI fallibilities. We add to the prior work by explicating who can be the verification worker and how to assess their wellbeing. We unpack why the need for such workers arose, why the frontline healthcare workers could not be the verification workers, and why the verification workers should be included while conducting human-centered evaluations of AI-based systems (deployed in public health).

2.4 Design-Based Implementation Research

Design-Based Implementation Research (DBIR) is a framework developed by researchers and practitioners in the field of education to design effective, scalable, and sustainable initiatives in the field of education [28]. As explained on a dedicated website[4]: *"It is an emerging method of relating research and practice that is collaborative, iterative, and grounded in systematic inquiry."*

Prior work in HCI has made a case for how DBIR can effectively approach HCI researchers and practitioners to co-create real-world impact [55]. This has also been echoed in other studies [41, 56]. Our study applies the four principles laid out by DBIR, which involve focusing on problems from multiple stakeholder perspectives, iterative and collaborative design, learning and implementation, and developing capacity for sustained change. We apply DBIR in the context of an AI-based intervention in public health.

2.5 The Capability Approach

The capability approach is a theoretical framework that measures human wellbeing by evaluating if a person has "substantive freedoms—the capabilities—to choose a life one has a reason to value" [94]. The approach was pioneered by economist-philosopher Amartya Sen and philosopher Martha Nussbaum [87]. Sen, in particular, argues that a person's wellbeing evaluation should assess if an individual has a substantive opportunity to achieve things they have a reason to value. It is an inherently pluralist approach where individuals may value different things, and the evaluation considers each person as an end—assessing if each individual has substantive (economic, social, and political) freedom to do and achieve things they value. The approach has been used in the context of human development [77] and broadens its evaluation focus from relatively narrow metrics of Gross Domestic Product (GDP) [58] to human capabilities—a person's substantive freedom to do and achieve things they want in life and have a reason to value.

We leverage this theoretical framework as our analytical lens. We build on prior works arguing for leveraging this approach to evaluate technology outcomes [74, 96] and designing technology that helps expand a person's capabilities [93]. We highlight the

[4] http://learndbir.org/

substantive freedom available to choose between different AI models in the context of public health. We also shed light on what our participants wanted to achieve through technology and argue to expand the focus of human-centered evaluations of AI.

3 BACKGROUND

Our study is set in the context of HealthNGO's Remote Engagement Service's (RES) India operations. We provide an overview of the work done by HealthNGO, introduce the Care Companion Program (CCP) and RES, and describe the multiple stakeholders involved in running the service, all critical to understanding our research.

3.1 Overview of CCP and RES

Established in 2014, HealthNGO aims to enhance health literacy among patients and families of individuals from underserved communities in India, Bangladesh, and Indonesia. Partnering with the local and federal government(s), HealthNGO provides patient and caregiver-centric health education sessions in public hospitals. These sessions are provided across several condition areas, including maternal and child health (MCH), cardiac care, tuberculosis care, and Non-Communicable Diseases (NCD) care. The sessions are part of the NGO's flagship Care Companion Program (CCP). In India, which is also the context of this study, HealthNGO provides CCP sessions across nine states in partnership with the state governments across 10,000+ healthcare facilities.

The CCP sessions are in-hospital health education sessions conducted by hospital nurses for the patients and their caregivers present in waiting areas of the hospitals. At the end of the session, the nurse presents a mobile number to the patients and their families, which they can call to enroll in a remote health education service offered by HealthNGO. This service, called the Remote Engagement Service (RES), was started in 2019 and is provided over WhatsApp[5] and Interactive Voice Response (IVR). RES helps HealthNGO to remain involved with community members after they leave the hospital. HealthNGO sends health-related messages to its registered members, i.e., the care recipients, over the IVR-based call or a WhatsApp-based message. The care recipients can also send messages and queries using HealthNGO's WhatsApp number. At the time of the study, approximately 1 million care recipients were subscribed to RES.

3.2 Medical Support Executives and Tele-Trainers

The queries posed by care recipients are routed to nurses employed within HealthNGO, called Medical Support Executives (MSEs). We note that the MSEs differ from the public hospital nurses conducting the training sessions. The MSEs are full-time HealthNGO employees who perform health information work within RES. The MSEs hold a nursing degree and/or General Nursing and Midwifery (GNM) diploma. There is a team of 16 MSEs in India, and their main task is to answer health-related queries from care recipients via WhatsApp. The MSE team can currently answer questions in six different languages: English, Hindi, Kannada, Marathi, Punjabi, and Telugu. To answer recipient queries, MSEs make use of cloud-based software that helps them view all messages asked on HealthNGO's WhatsApp number. Figure 1 shows MSEs' workflow in RES. Each MSE has a *bucket* allocated to them, and the care recipients' messages are automatically assigned to those buckets in a round-robin fashion [84]—a few messages (like single-letter messages) are discarded by a heuristic function. The MSEs refer to the number of messages assigned to them as message load or simply *load*. Once the MSEs see a message in their bucket, they either discard it, i.e., *close* the message as being *non-medical* and not requiring further attention or create a *ticket* for a medical message, i.e., health-related query. The MSEs and their managers track the number of tickets created and escalated daily. This is one of the metrics used by HealthNGO to keep track of their impact within the community.

The MSE may then respond to the medical message by answering the query by searching through a Frequently Asked Questions (FAQ) dataset, probing the care recipient to gather further details, or deferring the response and escalating it to the doctors working with HealthNGO. In the final case, the MSEs first translate the message into English, then escalate it to the doctor, and then translate it back to the recipient's native language before responding to the recipient. The MSEs work 6 days a week, with the entire team working on the weekdays and half-team working over the weekends. They work each day for 8 hours from 9 am to 6 pm with a lunch break from 2 pm to 3 pm. All the MSEs work remotely and are provided with a laptop by the HealthNGO.

Along with the MSE team, HealthNGO has a team of 40 Tele-Trainers (TTs) whose main task is to provide phone-based health education to select care recipients. Each day, the TTs get a list of 20-25 care recipients they need to call and provide information on

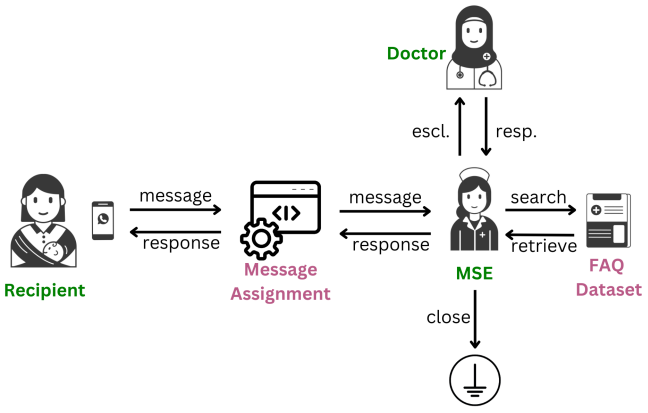


Figure 1: MSEs' workflow in RES.
The messages from the care recipients are assigned to the MSEs' buckets in a round-robin manner. Some messages (like single-letter messages) are auto-closed based on heuristics, meaning that all messages are not assigned to the MSEs. The MSEs respond by either looking up the answer to the health-related query in an FAQ dataset, asking HealthNGO's doctor for an answer, asking the recipients further questions, or closing messages that are not health-related.
(Abbreviations: escl. means escalate, resp. means response.)

[5]WhatsApp is an instant messaging and voice-over-IP (Internet Protocol) service owned by Meta Platforms Inc. (https://about.meta.com/).

health-related topics like pregnancy, infant care, and family planning. Out of the team of 40 TTs, two were moved to the MSE team in 2023 to support the MSEs with handling non-medical messages received in MSEs' buckets. These 2 TTs, who are critical to our study, reviewed the buckets of all the MSEs to identify and close the non-medical messages they could find. They sifted through the buckets of all the MSEs and worked in parallel to help abate their workload.

At the time of the study, the MSEs received around 10,000 messages. The total incoming message volume nearly tripled in the last year. Approximately 80% of the total incoming messages were non-medical, and the MSEs created 1,000 tickets daily. The MSEs aim to respond to the recipients in around a day or two based on the message load. The two TTs were onboarded to the MSE team to manage the high workload of non-medical messages. For future service scales and to cater to more recipients' health-related queries daily, the HealthNGO sought support through automation. They wanted to reduce the non-medical message workload on the MSEs and augment their ticket-creating capacity. We collaborated with HealthNGO to design an AI-based intervention with the aim of achieving this goal.

PID	Title	Exp.	Lang.	Gender	Part.
MSE1	MSE	3+	En,Hi,Pn	F	I+FGD
MSE2	MSE	1+	En,Hi,Kn	F	I+FGD
MSE3	MSE	2+	En,Hi,Kn	F	I
MSE4	MSE	<1	En,Te	F	I
MSE5	MSE	1+	En,Hi,Pn	F	I
MSE6	MSE	<1	En,Kn	F	I
MSE7	MSE	1+	En,Hi,Pn	F	I
MSE8	MSE	1+	En,Hi,Mr	M	I
MSE9	MSE	1+	En,Hi	F	I
MSE10	MSE	<1	En,Hi,Mr	F	I
MSE11	MSE	-	En,Hi,Kn,Te	F	FGD
TT1	TT	2+	All six	F	I
TT2	TT	2+	All six	F	I
M1	RES Mgr.	5+	-	F	I+FGD
D1	Dir.	7+	-	M	I
DR1	Dr.	6+	-	F	I
PM1	PM	6+	-	F	I

Table 1: Information about our study participants. 'PID' refers to the Participant ID. 'Title' is the official title of the participant at the HealthNGO. 'Exp.' is the number of years the participant has worked with the HealthNGO. 'Lang.' refers to the languages the MSE or TT responds to in their health information work. 'Gender' is the self-identified gender of the participant. 'Part.' means the type of participation in our study. 'I' refers to the participants we interviewed, and 'FGD' refers to the participants with whom we had our focus group discussion.
(Abbreviations: Mgr. means Manager, Dir. means Director, Dr. means Doctor, PM means Product Manager, En means English, Hi means Hindi, Kn means Kannada, Mr means Marathi, Pn means Punjabi, Te means Telugu.)

4 METHODS

We investigated how an AI-based intervention could reduce the workload on MSEs and augment their ticket-creating capacity. We adopted the Design-Based Implementation Research (DBIR) approach [28] to inform practice and contribute to the HCI scholarship. We studied the problems from a multi-stakeholder perspective and developed and iteratively improved an AI-based intervention. We conducted a 3-phase deployment where each phase informed the design of the subsequent phase. We critically examined our intervention through a mixed-methods analysis. The research was approved by Institutional Review Boards in India and the United States, where our multi-regional team was located.

4.1 Data Collection

We worked closely with the HealthNGO for 9 months, from January 2024 to August 2024. Our study began at a time when HealthNGO was exploring the potential of leveraging AI to support its scaling-up efforts. We supported this inquiry and collected data to understand the reason for this requirement and helped with the design and evaluation of an AI-based intervention. Our data collection included observations, interviews, a focus group discussion, intervention usage logs, and chat logs of MSEs with care recipients. We audio-recorded semi-structured interviews after seeking participant consent, except for three interviews in which the participants preferred not to be recorded. We anonymized all collected data and took care not to have identifiable information in the audio recording. The first author collected all the data in English, and the researcher ensured that all the interview participants were comfortable speaking in English when conducting interviews and the focus group discussion.

4.1.1 Observations. We carried out our observation through a mix of a 7-week in-person visit to HealthNGO's office and virtual collaborations. When remote, we caught up with the engineering team—developing the AI-based intervention—daily on a 30-minute call on Google Meet[6]. Besides this, there was a weekly hour-long debrief session on Google Meet with product managers to delineate the progress in designing and developing the AI-based intervention, discuss the challenges faced, and brainstorm the path forward. We maintained extensive handwritten and digital notes during our observations. The first author also attended 1 CCP session in person in a district hospital in Haryana, India, to better situate the importance of RES.

Starting in June 2024, along with observing the design and development efforts of the engineering team, the first author started engaging with the MSEs. The author joined their weekly hour-long calls on Google Meet and engaged with their communication channel on Slack[7]. This engagement helped us understand the challenges faced by the MSEs and establish a relationship with the team before conducting individual interviews. We maintained extensive handwritten and digital notes during our observations, which we used for our analysis.

[6]A video communication software developed by Alphabet Inc. (https://meet.google.com/landing).
[7]A team communication software developed by Slack Technologies, LLC (https://slack.com/).

4.1.2 Interviews. We conducted a total of 16 semi-structured interviews, including 10 MSEs, 2 TTs, and one each with the MSE response manager, product manager, doctor, and executive director. Half of these interviews were pre-deployment, and the other half post-deployment of the intervention. The interviews with the MSEs and TTs unpacked their background, career aspirations, challenges faced in their work, and their current awareness of and engagement with AI. In our post-deployment interviews, to mitigate the priming effect, we did not specifically probe the MSEs and TTs regarding the impact of AI-based intervention and just evaluated if there was a change in the nature of challenges expressed by the participants.

With other stakeholders, we inquired about their views on RES, its prevailing challenges, and how they felt AI could support their efforts in scaling. Table 1 provides more details about the study participants. We conducted our interviews over Zoom[8] except for one participant who spoke with us on Google Meet due to technical challenges with Zoom.

4.1.3 Design Probe. We engaged with the engineering team to design and develop an AI-based intervention based on our observations and pre-deployment interviews. This intervention helped act as a design probe [30], which was iteratively enhanced based on collective feedback from the stakeholders and its performance. The AI-based intervention was specifically designed to filter out the non-medical messages coming into the MSEs' buckets. To this end, we developed an AI-based intent classification model that could predict the intent of the incoming message and filter it out if it was a non-medical message. We started with the use of the GPT-4 LLM [1] as our classifier but shifted to GPT-4o [38] when the latter was released in May 2024.

We designed and developed this intervention over 4 months and deployed it in phases in the next 5 months. Figure 2 shows the different phases of our deployment. We carried out the deployment in 3 phases, each phase spanning a month on average. We took inspiration from prior works [49] and chose to deploy the intervention in phases to reduce the unanticipated outcomes from our intervention's deployment in the high-stakes domain. The multi-phase deployment also helped us assess MSEs' experience with the intervention. We iteratively analyzed the data from each phase to inform the subsequent phase's system design. The first phase was deployed entirely on the back end to collect data on model performance in production. We interviewed the MSEs and analyzed the model performance after this phase, which informed our next phase. Phase 2 involved making the AI-based intent classification visible as tags on the software interface used by the MSEs. We asked the MSEs to highlight the incorrect tags and continue with their usual operations. This phase helped us get data on MSE-annotated incorrect labels and assess MSEs' comfort in leveraging the intervention. We did not filter out any message based on the AI predictions. Similar to Phase 1, we conducted interviews and analyzed model performance after this phase. Finally, our Phase 3 deployment included filtering out the non-medical messages predicted by the AI system and assigning only the medical messages to the MSEs. We divided Phase 3 into three sub-phases where sub-phase 1 involved rolling out the system to 7 MSEs, sub-phase 2 involved rolling out the system to the other 8 MSEs, and sub-phase 3 involved rolling out the system to all the 15 MSEs (1 MSE was on maternity leave during the course of deployment). The sub-phase-based system deployment helped us evaluate our study's critical aspects, like the distribution of load between the different workers involved in our intervention's design. We revised our system iteratively till the full deployment. Based on the Phase 2 analysis, we added another set of workers, i.e., TTs, to the RES workflow, and the sub-phase design helped us understand the appropriate load that the TTs could handle. The sub-phases also helped us rotate the MSEs introduced to Phase 3 so that all the MSEs were acquainted with the intervention before our full deployment.

[8]A video conferencing software developed by Zoom Communications, Inc.

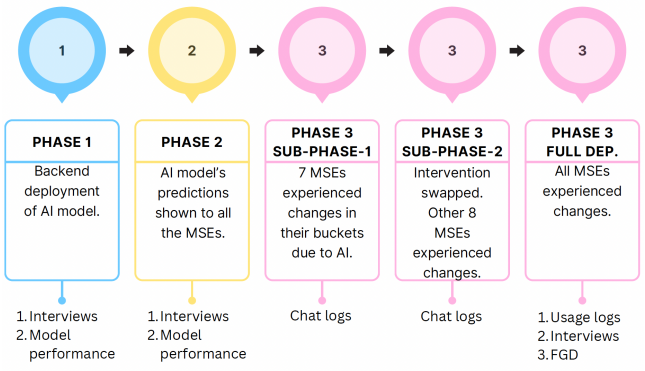


Figure 2: Process flow for our multi-phase deployment. In the first phase, the AI model predicted the intent of all incoming messages on RES. In the second phase, we showed the AI predictions to all the MSEs on the interface they used to read recipients' messages. We did not filter out any message based on AI prediction. In Phase 3, we started filtering out messages based on the AI model's predictions. In the first sub-phase of Phase 3, seven MSEs experienced changes in the workflow. We revised our system based on the analysis of the chat logs. The rest eight MSEs experienced these changes in sub-Phase 2 in Phase 3. Full Deployment refers to the stage where all the MSEs experienced messages being filtered out due to the AI model's predictions. We analyzed the usage logs, conducted interviews, and an FGD post-deployment. (Abbreviations: dep. means deployment.)

4.1.4 Chat Logs. At the end of each phase, we reviewed the chat logs of the AI model's incorrect predictions of a medical message. Moreover, during the Phase 3 deployment, we worked closely with the MSEs and TTs to identify incorrect predictions daily. We maintained a list of incorrect predictions in an online document and reviewed them daily to identify possible reasons for inaccuracy.

4.1.5 Focus Group Discussion. We conducted an hour-long Focus Group Discussion (FGD) with 3 MSEs and 1 Response Manager as our participants. We wanted to investigate the impact of the AI-based intervention post-deployment. We conducted this FGD around 40 days after the full Phase 3 deployment when the deployment stabilized. The FGD involved 2 MSEs whom we had interviewed pre-deployment and one other MSE whom we had not spoken with earlier.

4.2 Data Analysis

We followed Charmaz's guidance [19] for open coding of our observation notes, interviews, and FGD transcriptions. Through our pre-deployment coding exercise of the interviews and observation notes, we constructed themes such as, "MSEs finding joy at work", "stress in meeting targets", and "having uncountable messages". Next, we analyzed our design probe's usage logs iteratively after each phase. The analysis of the chat logs after Phase 1 deployment informed of errors like dates consistently misclassified by the AI model. We created a new *date intent* to resolve this error and considered it in the medical category. We evaluated the results of Phase 2 deployment using the F1 score [72] and false negative rate [72]—we defined messages predicted as *medical* in the positive class. The quantitative evaluation helped build confidence in the system's performance on the production data. This test production data consisted of 59K recipient messages across all six languages, with Hindi having the highest representation (23K) followed by Telugu (22K), Punjabi (7K), Marathi (3K), Kannada (3K), and English (1K). Finally, the continuous chat log analysis of Phase 3 deployment helped us catch errors in incorrectly parsing audio and image messages and misclassifying *yes* or *no* responses to MSE-initiated inquiries, which should have been considered medical due to the relationship with conversation history, among others technical bugs.

The authors met periodically throughout the coding process to discuss the codes and analyses from iterative deployments. We performed a second level of open-coding exercise post-deployment and constructed themes such as, "less leisure," "reminiscing learning," and "creating extensive tickets." This helped us sense a shortcoming between what the MSEs were looking to achieve through the AI-based intervention and what our design probe's usage logs showed. We drew on Sen's Capability Approach [94] as our analytical lens to assess the impact of our intervention on supporting MSEs' well-being. Analyzing AI's role in helping expand MSEs' opportunities to achieve their aspirations helped us identify a sociotechnical gap between what MSEs aspired to achieve through the intervention and how we measured success based on quantitative load reduction and ticket counts. We performed a third level of focused coding exercise, analyzing our entire data together, and constructed three broad themes: "Social Determinants of AI Model Choice," "Domain-Sensitive Human Infrastructure," and "Limitations of (Only) Assessing Human Capital."

4.3 Positionality

All authors are of Indian origin. We, collectively, have several years of experience working with frontline healthcare workers in the Global South. We have conducted studies in the context of Indian public health infrastructure for the past several years. Through our research and practice, we attempt to immerse ourselves in the lives of care workers. We recognize and represent their assets and work in solidarity to mitigate their challenges. We are committed to just and caring futures for all.

5 FINDINGS

We found a gap between what the MSEs wanted to achieve through the AI-based intervention and how our human-centered evaluation metrics measured the intervention's success. This gap aligned with prior recognition of a sociotechnical gap between social requirements from technology and its technical feasibility [2]. We examined this gap by studying the "technical feasibility" of AI and "social requirements" through AI. We unpack why we chose a specific AI model (that determined technical feasibility), whose social requirements must be considered while conducting human-centered evaluations, and what those social requirements were.

5.1 Why LLMs: Social Determinants of AI Model Choice

The HealthNGO envisaged using AI to reduce the workload of MSEs and scale RES. The choice of technology was based on HealthNGO's awareness of the technology, evidence of its usage by peer organizations, and alignment with RES's future goals. Our specific AI model's choice, i.e., GPT-4o LLM, was based on a number of social and economic factors. There was substantive social and economic freedom to choose LLMs compared to other (smaller or regional) AI models. Our choice of specific LLM was based on access to GPT-4o (through ChatGPT[9]), financial freedom to leverage GPT-4o, and availability of (technical and human) resources to develop and deploy GPT-4o. These broader social and economic factors shaped model choice on top of the model's performance on quantitative evaluation benchmarks.

5.1.1 Access to LLMs. Public access to LLM-based chatbots through web and/or chat-based interfaces (like ChatGPT and Meta AI[10]) played a crucial role in determining the choice of an AI model. It helped increase awareness of technology's abilities, assess its usage in daily workflow, and identify potential use cases most suitable for such AI models. MSE4 described her use of ChatGPT as:

> *"Everywhere now AI [is] there ... [I use] ChatGPT [on a] daily basis ... for like translations and all. Like, if [I have] big conversations from Telugu to English [that] I need to translate, ... and I can't sit and type everything ... [I use] ChatGPT ... Almost [all] translation[s] [are] correct"* (MSE4).

The access to and use of ChatGPT helped MSE4 understand its utility in her work. While MSE4's expression of AI's omnipresence reflected a consequence of a global narrative—through geopolitical and academic institutions and multinational corporations—perpetuating the AI hype [67], her use of the tool helped her assess the chatbot's performance. She showed us how she accessed the interface and corrected the inaccuracies in model-generated translations. Similarly, other MSEs (who had interacted with AI-based systems) narrated how they appropriated such chatbots for different tasks in their work, such as embellishing their responses to community members (MSE1) or acquiring health and non-health-related information (MSE3).

This straightforward access to LLM-based chatbot shaped perceptions around potential future use cases of AI in RES. Participants narrated various ways AI could abate scale-induced challenges in RES. The ideas presented by the participants were shaped by their

[9] ChatGPT is a generative AI chatbot that was made openly available for public use in 2022. The chatbot provided limited and free access to GPT-4o LLM through a web application (https://chatgpt.com/) at the time of writing this paper.
[10] Meta AI is generative AI chatbot available through WhatsApp. This chatbot became publicly available during the course of our study.

experience with LLMs and their understanding of the prevailing limitations of the service. AI usage by peer public health organizations also used LLMs, which played a critical role in exploring avenues for the future use of AI. Since a majority of these use cases were shaped by personal use and assessment of LLM-based chatbots, LLMs were the most suitable AI model choice for the envisioned tasks. DR1, a doctor in the HealthNGO, narrated her perception of AI utility as:

> *"AI can significantly support MSEs by enhancing their ability to ask more insightful, direct, and targeted follow-up questions, enabling a deeper understanding of the user's [care recipient's] problem. ... AI can also assist by selecting the most 'relevant' [emphasis added] responses and can be further trained to respond more 'empathetically' [emphasis added], improving the overall quality of care"* (DR1).

DR1, who had been using ChatGPT in her work, highlighted how AI could support the MSEs in triaging by asking follow-up questions to the care recipients' inquiries. She emphasized providing relevant and empathetic responses to the recipients, which, we found, were challenges experienced by the HealthNGO in RES. Most use cases were suitable for a chatbot-like technology, for which LLMs were the more suitable AI model type.

The HealthNGO, however, assessed the envisioned use cases through a critical lens. Having experienced inaccuracies in their use of LLM-based chatbots (due to hallucinations [42]), our participants expressed skepticism—in contrast to prior work making a case for "AI authority" in similar application contexts [43]—towards the ingenuous acceptance and use of model output. MSE5 expressed, *"we can't believe that AI is always correct because sometimes, due to [a] language barrier or something, [even if] one sentence [is] also [wrong] ... or right, [we] can't [let AI respond to recipients]."* She narrated how she and her colleagues recognized the importance of expert oversight before responding to recipients' queries. HealthNGO's director (D1) expressed cautious optimism towards using AI. While having a concrete AI vision document, he expressed placing central importance on HealthNGO's mission of enhancing health outcomes among care recipients. He saw AI as a tool to help achieve HealthNGO's goal and encouraged experimenting with AI and evaluating its impact. Along with use-case appropriateness, the model choice for these experiments was determined through the availability of human and technical resources to experiment with specific AI models, which we describe next.

5.1.2 Implementation Freedom. The first AI use case chosen by the HealthNGO was to support MSEs in reducing their workload. In our pre-deployment interviews, the MSEs expressed skimming through numerous recipient messages that required less attention from the MSEs. This delayed response to recipients who required higher attention. MSE1 expressed:

> *"We have to eliminate all the non-medical [messages] because our more [sic] time is going to close non-medical [messages] ... There [are] a lot of non-medical [messages], so we are not able to focus on medical [messages]"* (MSE1).

She wanted her message bucket to contain only medical messages, i.e., messages from recipients who required health-related support. Her bucket currently contained copious non-medical messages that included acknowledgments (like "thank you," "okay," or "done"), greetings (like "Hello, good morning, Ma'am," "Sir, reply"), and spam messages (automated replies, offensive terms), among others. Other MSEs shared similar experiences and expressed that the increasing service scale exacerbated this challenge.

We designed our AI-based intervention to ameliorate this challenge. We found that the relative ease of developing an LLM-based classifier affected our choice of the AI model. We discussed using traditional machine learning (ML) classification models (like SVM [35]) on top of small multilingual language models (like mBERT [24]) but delayed implementing such a technical system due to its perceived complexity. While we anticipated that these relatively smaller models might be more suitable for our task and may have a lower financial cost, they required the creation of a model training dataset from scratch and relatively higher familiarity with traditional machine learning models, i.e., more human resources. On the other hand, LLMs were available through Application Programming Interface (API) calls and required less training data during development. The dataset used in our model was developed by ~1 MSE per language and comprised of ~114 examples per language across six languages. Our AI model—based on Retrieval Augmented Generation [60]—required two engineers and the first author to develop the first working version of our AI-based intervention within a few days of receiving the data.

Moving from model development to system deployment, the open-source LLM engineering systems (like langfuse[11]) helped with setting up a deployment and monitoring pipeline. We followed the recommended practices in the operationalization of LLMs [98] to help with version controlling of our prompts and conducting periodic evaluations of the model. We (in coordination with MSEs) also made policy decisions regarding what messages should show up in MSEs' buckets and how to handle hallucinations. This helped us slowly build a sociotechnical infrastructure around the use of LLMs as our choice of AI model—only a part of which could be transferred to alternative AI models.

5.1.3 Financial Freedom. The third social factor that determined the choice of a specific AI model was the financial cost of implementing or experimenting with the model. The HealthNGO won an impact grant that awarded monetary credits to use LLMs, which helped with the financial freedom to experiment with the model in the intended use case(s). Over the course of the study, HealthNGO won two more such awards for working on the proposed use cases through LLMs. This provided substantive financial freedom to use a specific LLM type against other LLMs (like regional LLMs).

We compared different LLM types offered by the technology organization that offered the credits, i.e., GPT models by OpenAI[12]. We evaluated the models we developed using F1 score [72], false negative rate (FNR) [72], and the cost of implementing the model—an evaluation criterion found to yield better AI-mediated system design [45]. During the course of our study, GPT-4o was launched, and we shifted from GPT-4 to GPT-4o in Phase 3. While there was

[11] https://langfuse.com/

[12] OpenAI is an AI research and deployment company (https://openai.com/about/).

a 28% performance decay (based on FNR) in moving from GPT-4 to GPT-4o, we found a 50% cost reduction in our intervention (after making the model switch to GPT-4o). In making the switch, the average FNR—across languages—increased from 0.60 to 0.83, and the cost decreased from $10 to $5 per day.

Table 2 shows the performance of the GPT-4o model as used in Phase 3. There was around 1% error in the final model based on FNR as our evaluation metric. While this was a relatively small error rate, the HealthNGO was still unsure of an unchecked system deployment at scale—which could lead to inequitable health outcomes. This helped us examine the extra considerations required when deploying LLMs in high-stakes domains. Our considerations and workaround for AI fallibility are described next.

Language	F1 Score	FNR
English	97.30	0.18
Hindi	95.87	0.67
Punjabi	95.01	1.36
Marathi	94.93	0.75
Telugu	94.59	0.99
Kannada	94.06	1.03

Table 2: Performance of the GPT-4o-based intent recognition model as used in the phase 3 deployment.

5.2 Whose Evaluation: Domain-Sensitive Human Infrastructure

We reduced the AI model's false negative rate to around 1% through iterative improvements, as shown in Table 2. Deployment of this model still meant faltering on 100 messages daily, which was unacceptable for HealthNGO—operating in a high-stakes domain and committed to providing equitable care to the recipients. The high-stakes nature of our deployment guided us to look for ways to balance reduction in MSEs' workload with human oversight of recipients' messages. We now describe how we mitigated AI fallibility in our intervention and its impact on the MSEs' workload.

5.2.1 Adding Verification Workers in RES. The requirement of human oversight to mitigate AI fallibility became axiomatic after Phase 2 deployment. Despite continuously collecting and correcting model errors (by improving our dataset) and enhancing our prompting strategy, we could not reduce the error to less than 1% on average. Subsequently, we changed our inquiry from asking *when* to deploy—which meant reaching the least possible error rate—to *how* to deploy—which meant looking at ways to incorporate human oversight. We deliberated on continuing the Phase 2 design strategy that showed the AI classification tags without filtering out any message to the MSEs. This design is helpful in content moderation [50], and the MSEs also found them useful in their work. However, the MSEs wanted to reduce sifting through non-medical messages in their buckets. To help lessen the workload for the MSEs and accommodate human oversight of AI predictions, we included another set of workers, Tele-Trainers (TTs), in RES.

In 2023, the HealthNGO asked two TTs to shift from their usual tele-training work to health information work and support MSEs in addressing recipients' non-medical messages. To make this transition, TTs completed a week-long training followed by a week-long probationary period, where each training and observation session lasted for around three hours. In our conversation with the TTs, both mentioned being able to pick up the work of discriminating between medical and non-medical messages and addressing the latter with relative ease. They saw this shift as becoming a part of the MSE team and as part of their career progression. Their work-life balance improved, they started learning new languages (to address messages from unfamiliar languages), and they sought mentorship from the MSEs to enhance their technical skills. We leveraged their acquired assets in our system design to mitigate AI fallibility.

Figure 3 shows the design of our intervention. Our system design involved a minor change in TTs' workflow. Before our intervention, the TTs reviewed the buckets assigned to MSEs to address non-medical messages. We now created individual buckets for the TTs and assigned all the non-medical messages predicted by the AI model to the TTs (at random). The TTs were asked to escalate the messages they felt were medical to the MSEs. The TTs got comfortable with the new workflow within a day and particularly appreciated the creation of personal buckets. TT1 mentioned:

> *"Process of tagging is now very streamlined. [The] creation of [a] separate bucket is very helpful. [I] used to search for non-medical and medical [messages in MSEs' buckets] ... [Now] time is being saved ... Almost all [messages are] non-med[ical] in my bucket"* (TT1).

We found that our intervention provided auxiliary support for streamlining TTs' workflow and corroborated with prior work that has highlighted the assistance to data work through AI-based interventions [41]. TT1 also expressed that most of the messages in her bucket were non-medical, which helped her address them quickly without searching for them in MSEs' buckets. Both TTs

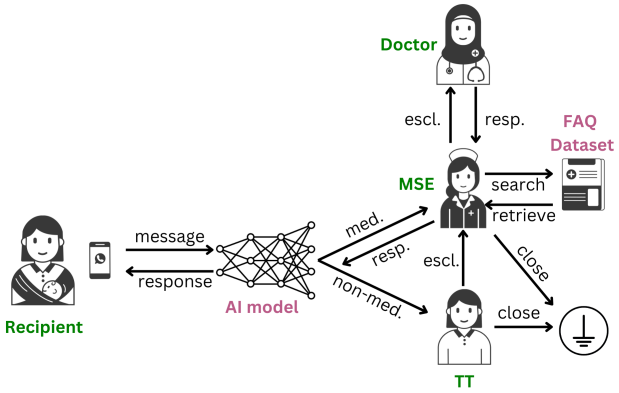


Figure 3: Revised RES workflow to mitigate AI fallibility. The messages predicted as "non-medical" by the AI model are directed to Tele-Trainers (TTs). The TTs can then either escalate the message to MSEs or close the message. The TTs and the MSEs closed a recipient's message when it was not considered a health-related query.
(Abbreviations: med. means medical, non-med. means non-medical, escl. means escalate, resp. means response.)

expressed seeing medical messages in their buckets, which they escalated to the MSEs. We evaluated the number of such errors in our Phase 3 deployment and assessed our system design's impact on both the MSEs and the TTs, which we present next.

5.2.2 Evaluating Impact on MSEs and TTs. We found that our AI-based intervention helped reduce the load on the MSEs and augment the MSEs' capacity to address the queries of a greater number of recipients, i.e., create more tickets. Figure 4 quantitatively illustrates our intervention's impact.

First, as shown in Figure 4a, the TTs smoothly transitioned to their new workflow and could catch the AI model's incorrect predictions. The initial high error rate was due to our policy of marking multimedia sent by recipients (like voice notes and images) as non-medical messages, which we realized were usually medical (and we corrected it after the first sub-phase). Our deployment stabilized after the end of our second sub-phase, i.e., around day 21, and the TTs could catch a few incorrect AI model predictions each day. We found that apart from catching incorrect AI predictions, the TTs could also catch some of the messages previously missed by MSEs (due to human error and high service load). In the initial days of the deployment, the TTs could carefully go through recipients' conversation histories and identify any unanswered messages that required the attention of the MSEs. Their feedback helped us understand the gaps in our deployment and enhanced the scrutiny of the recipients' queries.

Second, the number of messages assigned to MSEs gradually reduced during the course of our deployment. As shown in Figure 4b, our deployment started with around 60% of incoming recipients' messages being assigned to the MSEs, and this reduced to around 40% towards the end, and the MSEs confirmed this reduction in our post-deployment interviews. The message assignment load on the TTs, on the other hand, increased rapidly during the course of our deployment. We found that within a few days of ending the second sub-phase, the TTs started finding assigned messages as *"uncountable"* (TT2). The TTs expressed that the workload during sub-phase deployment—when the MSEs to TT ratio was around 4:1—was manageable but eventually became intractable (the HealthNGO added three more TTs in the MSE team based on the recommendation of this study). Here, we found that though our human-centered evaluation aimed to investigate the intervention's impact on MSEs, assessing the intervention's impact on TTs (verification workers) also became crucial.

Third, there was a veritable increase in the number of tickets created by the MSEs. Figure 4c shows the average number of tickets created per week during the deployment period. The reduction in the number of non-medical messages allowed MSEs to address the medical queries of a greater number of recipients and create more tickets. The MSEs surpassed their personal best twice during the course of our study and were valorized for creating record-breaking tickets. The expectations simultaneously increased from the MSEs as their work was now supported through AI. The MSEs themselves anticipated further augmentation of their abilities as expressed by MSE1, *"AI can help us so that instead of 1000 we are able to create 2000 tickets per day."* However, the augmentation of MSEs' ticket creation capacity had a limited impact in reducing their work. In our FGD that was conducted after the entire deployment period, we found that the MSEs still experienced a high workload with reasons beyond the occurrence of non-medical messages in their buckets. These reasons included the onboarding of new care recipients on RES and the distribution of the MSE team's capacity on other projects. The MSEs also performed invisible work that was not recorded in the ticket creation metric. MSE11 explained:

> *"If we are creating 100 tickets, it's not like 100 families only we have attended, it might be 250 families which we have attended, and we probed them, and a lot of back-end process will be there, and we have asked questions for them [from the doctors] ... If you [researchers and HealthNGO staff] are taking only ticket creation [for the MSEs' work evaluation], it will be unfair"* (MSE11).

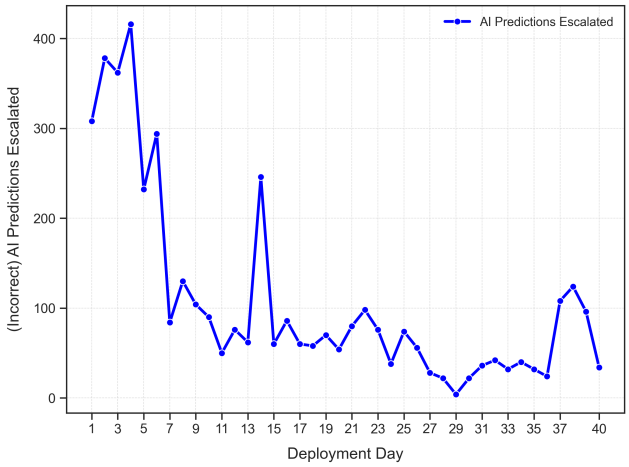
Here, MSE11 expressed how she followed up with care recipients without specifically creating new tickets for those conversations. She also mentioned performing invisible technical work like changing the status of recipients' delivery details in the database and following up on the concerns of the recipients, among others. MSE11 argued that it would be unfair for her work to be evaluated solely on the tickets she created. Our AI-based intervention helped reduce a part of the MSEs' overall workload but had a limited impact in reducing the MSEs' work burden as they started creating more tickets and continued with other invisible parts of their work. This complicated the goal of our AI-based intervention. Next, we shed light on a deeper examination of why the MSEs (and TTs) really wanted an AI-based intervention in their work.

5.3 What to Evaluate: Limitations of (Only) Assessing Human Capital

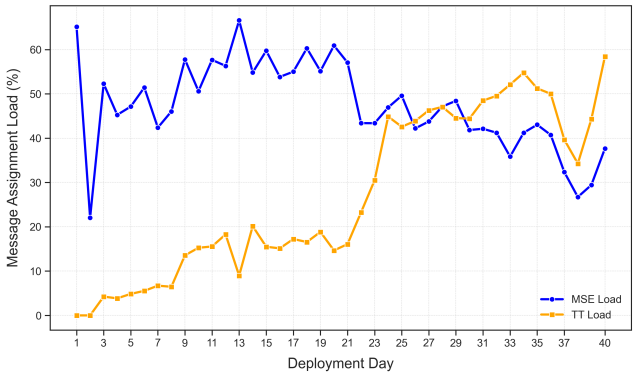
We re-examined what should have been the goal of our AI-based intervention, which helped us understand what we must evaluate as part of our human-centered evaluations. The MSEs and TTs wanted to leverage AI to achieve their broader aspirations and get opportunities to do the things they valued. A narrow focus on building human capital to help the workers become efficient and skilled, i.e., build their human capital through AI-based intervention, missed the broader things that the workers valued in their work. We shed light on these aspirations and highlight the things they valued. We unpack why the MSEs and TTs wanted to become efficient in their work, enhance their skills, and distribute their workload.

5.3.1 From Efficiency to Relationships. The MSEs wanted to leverage the AI-based intervention to support more recipients' health-related queries in less time. Achieving service scale is considered an act of caregiving among health information workers [47]. The MSEs wanted to achieve impact at scale and create more tickets because they valued the relationships they formed with the recipients in the process. MSE10 narrated her experience of relationship building as:

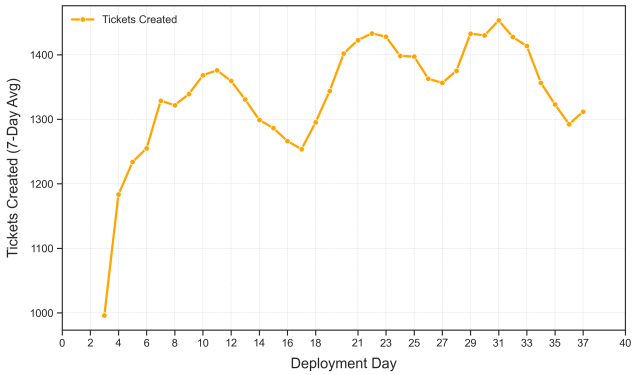
> *"There are [recipients'] families which turn back every day to us. Even if we cannot see who they are, from where they are, there is some connectivity ... when the patient [recipient] comes to us with new queries, I, in my mind, am like, 'I know this person.' Since she was pregnant, to [when] she gave birth to the baby, and*



(a) Incorrect AI predictions escalated by TTs to MSEs



(b) Message load assignment comparison between MSEs and TTs



(c) Weekly average of the number of tickets created by MSEs

Figure 4: Findings from 40-day deployment study of our AI-based intervention.
Day 1 is defined as the day when phase 3 was deployed. The sub-phase was completed on day 14, and sub-phase 2 was completed on day 21. The intervention was deployed for all MSEs on day 21. Figure 4a shows the count of messages that were predicted as non-medical by the AI model but were considered medical messages by the TTs and escalated to the MSEs. Figure 4b shows the percentage of RES's message load assigned to the MSEs and the TTs after the start of phase 3. Some recipient messages are auto-closed (like single-letter responses) in RES, meaning that the entire message load is not assigned to the MSEs or TTs. Figure 4c shows the weekly average of the number of tickets created by the MSEs. The average was taken in order to account for the reduced number of MSE staff working over the weekends.

then how her baby is and what all she's going through post-discharge ... [there] is an invisible bond" (MSE10).

MSE10 worked on a language that had a relatively smaller number of people signed up on RES (due to a new demographic for the HealthNGO). She received around 100 messages daily and handled the entire service for that language herself[13]. She could relate to the emotions of the recipient during the latter's pregnancy journey. MSE7 told us, *"when patients say that had an abortion in 4th or 5th month of pregnancy or that their baby died ... being a female, I can understand the pain."* These emotional bonds that the MSEs formed with the recipients were what MSE10 called an *"invisible bond."* The MSEs found purpose in their work (MSE4) and shared the recipient's pain when their support was found wanting by the recipients (MSE9).

The MSEs valued the feedback they received from the recipients. The positive feedback added meaning to their work. These feedback messages ranged from "thank you" messages to recipients making videos expressing their gratitude for MSEs' work and advocating for RES on social media (MSE5). The MSEs valued when the recipients, in turn, checked in on their wellbeing with messages like *"how are you Ma'am, did you [have] your tea?"* This complicated our understanding of what messages should be visible to the MSEs. While non-medical, such feedback messages added value and meaning to

[13]Another MSE supported at times service load increased.

MSEs' work, and our focus on reducing work for the MSEs' work and bolstering efficiency discarded such messages[14].

5.3.2 From Upskilling to Knowledge. The MSEs shared concerns regarding the future of their work with AI on the horizon [23]. They were concerned about how it might impact their job (MSE4) and how they could learn to use and work with AI. Upskilling and enhancing AI literacy is crucial with AI technology progress [48]. We found that there was also a deeper and broader reason why MSEs wanted to gain knowledge of AI. Beyond upskilling to get better at performing a limited number of tasks, the MSEs intrinsically valued the knowledge of AI. Without knowing specifically how AI skills could help them (MSE6), they anticipated that AI knowledge could be an avenue to achieve their aspirations. MSE2 told us that in her career, she aspires to:

> *"[keep] learning step by step and understanding the [medical] concept and understanding the family [recipient] ... [keep] updating our knowledge ... [medical] and non-medical ... I [am] waiting to know how AI can help us ... how we can use it ... so that we can learn and we can teach to others. So, if someone is asking me if you have heard about AI, 'Yeah! I have heard [about] it!' I feel that's a great thing"* (MSE2).

MSE2 found expanding her medical and technical knowledge intrinsically valuable. The medical knowledge helped her understand the concerns of recipients better, while she saw technical prowess as instrumental in acquiring medical knowledge. She wanted to learn how AI could help in different tasks at work. She wanted to learn how to use it to appropriate it for her own use cases. After learning, she wanted to teach it to others and found value in sharing her knowledge. Similarly, MSE7 correlated being introduced to AI (and other technology) and becoming *"hi-tech"* as the *"most rewarding thing"* she learned in her work.

The MSEs had started using AI like ChatGPT to learn words in new languages, for which they relied earlier on services like Google Translate[15]. A narrow scope of upskilling TTs to flag AI inaccuracies (Figure 4a) and MSEs to assess the AI-generated labels (in Phase 2) missed the broader reasons why the MSEs (and TTs) wanted to gain knowledge of AI. Acquiring skills helped them be better at specific tasks at work, but the knowledge of AI helped them evaluate how the tool might make sense for them and explore different ways they may want to use the technology.

5.3.3 From Division of Labor to Leisure. Both the MSEs and the TTs valued getting time for leisure in their work. The freedom to work remotely was an important reason why the workers chose to work with HealthNGO. It helped them spend time with their families and maintain a better work-life balance (as found in prior work [47]). They wanted a part of their work to get distributed to help reduce their workload. They wanted to use the freed-up time to do things that they found meaningful. TT2 expressed:

> *"Sometimes I also go to work with counseling tasks, ... I take leave and that time I [provide] voluntary service in [a] difficult case"* (TT2).

TT2 worked as a family counselor before joining the HealthNGO. She told us that she was *"proud of me [herself] sometimes"* for the work that she did before and took leaves from HealthNGO's work to go back and attend challenging cases. She found meaning and joy in the work she did and sought time off from her tele-training work to do another type of work. The division of work due to the AI-based intervention, on the other hand, had a negative impact on her freedom to take a leave. She was concerned about the effect of her absence on her colleague's (TT1's) workload. Our intervention should have accounted for the impediments imposed on the TTs when distributing the work away from the MSEs.

A sole focus on the distribution of work away from the MSEs did not mean that the MSEs got more leisure time. As the ticket creation process became easier, (intrinsic and extrinsic) expectations increased, and they created more tickets. The MSEs valued leisure time, not just time away from work. MSE5 told us, *"I will go through other projects after [message] load decreases."* She had experience participating with other teams within the HealthNGO (like the program service and design teams) and helping them with their work when her workload was less. In her leisure time, she wanted to continue expanding her knowledge and working with other teams on tasks unrelated to health information work. Along with becoming specialized in her work—argued as one of the merits of the division of labor [26]—she wanted to explore her interests and expand her skill sets by collaborating with other teams.

Our intervention helped distribute work between the MSEs and TTs, but the division of labor negatively impacted one set of workers while having a limited positive impact on the other set in terms of getting leisure time during work to do things that they found interesting. Building on Sen's argument [95], we argue that developing human capital to work more efficiently through AI does not imply that the workers are able to achieve and do things that they value. Our narrow evaluation focus on goals like achieving efficiency, upskilling, and distributing work among workers may miss broader goals of developing relationships, gaining knowledge, and getting leisure. We now discuss these points in detail.

6 DISCUSSION

We discuss the implications of our findings vis-à-vis HCI scholarship. We argue for expanding the scope of human-centered evaluation of AI by considering the social factors determining AI model choice, the role of our high-stakes domain, and accounting for worker aspirations. We reflect on these implications and present the multiple dimensions that need to be considered while conducting human-centered evaluations of AI.

6.1 Towards Informed AI Model Choice

We found that on top of our AI-based intervention's performance on evaluation metrics, social factors determined GPT-4o LLM's choice in our implementation. The HealthNGO's access to GPT-4o and substantive (implementation and economic) freedom to leverage an LLM in our intervention led to our model choice. Our study adds nuance to conversations within HCI deliberating the use of LLMs [6, 54] and arguing for adding "solid justification on why to apply LLMs" [5]. We argue that highlighting the social factors that made

[14]The AI model was revised after phase 3 and a separate feedback intent was created, which was directed to the MSEs.
[15]https://translate.google.com/

leveraging an LLM suitable for conducting research in HCI could be a valid justification for model choice.

We further argue that the reduction in social and financial barriers to leveraging AI models is not equitable. We found that HealthNGO had greater freedom to choose between the models offered by AI startups like OpenAI and Cohere[16] than to develop and deploy a traditional ML model like SVM [35]. We did not attempt training and evaluating the performance of a traditional ML model because of the high demand for human and financial resources. This is at odds with the "AI democratization" calls [92] as different AI models did not have an equal opportunity to be chosen by HealthNGO. Along with easier development, the high visibility of specific LLMs (like ChatGPT) narrowed the use cases of AI applications such that language models may be the most suitable technology choice.

Next, we argue that HCI research can play a crucial role in mitigating this inequity. First, designing publicly accessible interfaces that could highlight the abilities of traditional or smaller machine learning models on specific use cases may enhance its acceptance for similar use cases. Such interfaces may help establish the performative abilities of traditional and smaller ML models among individuals with less AI experience. Second, advancing the research on interaction systems that help develop traditional and smaller AI models with less implementation effort (like AutoML [39]) may help organizations with less technical implementation resources to develop smaller and use-case-specific AI models.

Reducing barriers to traditional and smaller ML models may still be unable to match the hype around LLMs [67] or may fall short—in performance—in comparison to LLMs. In such cases, we argue that it is important to make visible all the costs (and additional benefits) of using LLMs. These costs go beyond social costs (like highlighting the possibility of disinformation and toxicity in outputs) to include factors like LLMs' long-term financial and environmental costs. The HealthNGO currently had free credits to use the model API, but hoping to sustain the free credits may be less viable within capitalist structures [97]. Here, we add to calls from researchers arguing for LLMs to have energy ratings that could help technology users get information regarding the environmental costs of LLMs [65]. Similar to energy ratings, accounting for financial sustenance may help similar resource-constrained organizations to make an informed choice—also reducing technical debt [12] to migrate after credit expiry. Accounting for such social and financial barriers adds nuance and explanation to model choice—expanding the evaluative focus from *what* choices were taken [25] to also include *why* those choices were taken—and helps the technology users take an informed AI model choice.

6.2 Nurturing 'All' Humans in the Loop

We started our intervention to reduce the workload on the MSEs. Along the way, we found the need for TTs, i.e., verification workers, to mitigate AI fallibility. Prior studies leverage human verification workers like "overreaders" [11] without specifically accounting for them in human-centered evaluations. We add to such studies and argue for expanding the scope of who the humans in the loop are while conducting human-centered evaluations. We discuss three critical questions regarding designing with and for verification work(ers) in AI-based systems in public health.

First, we discuss *why* verification work is crucial. We argue that the existence of verification work in high-stakes domains like public health is likely to stay for a long time due to the intrinsic fallibility—owing to the probabilistic nature—of AI [42]. In recognition of this known limitation, the usage policies[17] of companies developing LLMs (like OpenAI) currently mandate verification by a "qualified professional" before leveraging the APIs to generate health-generated content [75]. Wolfe and Mitra proposed adding a "socio-technical verification dimension" in the design space of generative AI-based systems in high-stakes domain [112]. Their recommendation adds support to recognize the importance of human verification workers while highlighting concerns regarding the laborious nature of verification work and the possibility of devaluation of human labor. We add to this (and similar [21]) calls and argue that recognizing such work will be critical with technical advancements in AI. We call on HCI researchers to design futures that mitigate drudgery and undervaluation of verification work.

Second, recognizing that verification work will remain indispensable opens avenues to ask the next critical question, i.e., *who* will perform the verification work. Recent works have shown the potential of "LLM-as-a-Judge" [117] to perform verification work. Wolfe and Mitra also have "GenAI Verifier" on one end along the dimension of verification work [112]. We argue that the need for human verifiers will remain consequential due to the intrinsic fallibility of AI-based verifiers, as we discussed earlier. HCI research could play an important role in delineating the types of verification work that could arise and the associated skills that may need to be nurtured among the workers.

In our intervention, both the TTs had a Master of Science in Social Work, a required skill for the tele-training work they performed before joining the MSE team. They undertook light week-long training to join the MSE team and a week-long probation period. Both the TTs described relative ease in acquiring the skills for this work, and our deployment illustrated ease in catching and mitigating AI errors. While the formal education of both the TTs may have positively impacted the ease of acquiring critical thinking skills to assess AI's output, it may be possible to nurture these skills among individuals with less formal education. We found that the MSEs began critically assessing ChatGPT's performance after interacting with it and chose to dismiss its output based on their analysis of the output. Similarly, prior works have found how explanations could help develop similar critical assessment skills among frontline healthcare workers with lesser formal education [101]. It may be possible to build the capacity of individuals to perform verification work, and we could draw from studies presenting design recommendations to build different types of literacies—both AI-related, like incentivizing excellence in data work [90] and non-AI-related like mobile-based training for health education [116]. Recognizing and nurturing skills required for verification work may help create new jobs within the AI infrastructure.

Third, it is crucial to discuss *how* the verification work is designed. Concerns exist regarding the laborious nature of this work and possible undervaluation [112]. It is important for HCI scholars

[16] Cohere is another AI research company (https://cohere.com/about).

[17] Most recent policy update while writing this paper was on January 10, 2024.

to recognize the value of verification work. We draw from recommendations by HCI scholars working to enhance the value of data work in AI through enhancing meaningful public participation [14, 91] and/or structural reforms in the incentive structure of AI in research and industry [90]. We build on such recommendations and highlight that to mitigate drudgery in verification work, the number of workers could be increased and, in turn, decrease the average workload on individual workers. One way of achieving this is to draw recommendations from Boone et al. study of how citizen science projects attempt to ensure meaningful work for its volunteers [14]. Another way could be designing Human-AI collaboration for verification work that minimizes the need for verification [111]. Next, we recommend adopting design frameworks like "Data Feminism for AI" [53] that could help recognize the assets of the verification workers in the broader human infrastructure of AI. This framework embraces pluralism, recognizes the situated knowledge of individuals, recognizes the embedded power structures, and advocates for making labor visible. Designing AI-based systems based on similar frameworks could help recognize the role played by the verification workers, nurture their assets, and advocate for designing interventions that expand the verification workers' capabilities.

6.3 Evaluating Expansion of Capabilities Through AI

We found that on top of their work-related needs, the MSEs wanted to achieve broader aspirations at work. We argue that the goal of AI-based systems should be to enhance the significant opportunities available to users to achieve such broader aspirations. We draw from a similar call from the AI ethics scholarship [64] and discuss design considerations to expand human capabilities through AI.

First, we build on calls from HCI researchers to view AI model evaluation as "narrowing the socio-technical gap" [62] and add nuance to how the *social requirements* of individuals should be studied when conducting human-centered evaluations. We found that if we focused on what the MSEs and TTs *needed* to serve the recipients at scale, then our intervention could be considered a successful intervention. On the other hand, when we expanded our evaluative focus and studied the broader things the MSEs and TTs wanted to achieve in their work (and careers), we found that the intervention fell short of creating substantive opportunities to enable those aspirations. We recommend that along with adopting human aspirations-based design in HCI [57], researchers should focus on AI-based intervention's ability to create opportunities that can enable the users to achieve their aspirations when conducting human-centered evaluations.

Second, to create such opportunities, we recommend reconsidering the narrative of viewing AI interventions as a means to *augment* human abilities [22, 52]. We found that the reduction in MSEs' workload went hand-in-hand with the MSEs creating more tickets. The augmentation of their abilities impeded a reduction in their workload. We recommend viewing an AI intervention's ability to *reduce* workload, which may be a more effective narrative when conducting human-centered evaluations. If the goal is to reduce the work on the workers, then that may require additional considerations. For example, in our intervention, along with the AI-based intervention, a policy to create *protected time*—creating a time block during work hours for self-determined activities—for the MSEs and TTs may have helped reduce their stress and given them the opportunity to pursue their work-related aspirations. Prior works have shown how computer-assisted protected time can particularly benefit information workers [20]. We add that such a policy needs to be designed to take HealthNGO's aspirations to scale into account, and further considerations are needed to balance an individual's aspirations with organizational goals. We now discuss some of the broad dimensions of conducting such an evaluation.

6.4 Multidimensional Human-Centered Evaluations of AI

We draw on the multidimensional nature of the Human Development Index (HDI) [87] to propose a multidimensional approach to conduct human-centered evaluations of AI. HDI measures proxies for three central capabilities—health, education, and decent standard of living—to assess the development of individuals living in countries across the world [3, 102, 108]. We connect the implications of our findings with a recent blueprint of multi-layer foundational models—LLM-like AI models argued to underpin a variety of AI applications [13]—proposed by Suresh et al. [105]. The proposed blueprint aims to enhance the participation of human actors in the design of foundational models. It contains a *subfloor* and *surface* layer on top of the *foundation* layer containing the AI model. We connect our previous three discussion points with this multi-layer architecture and recommend three dimensions for conducting the human-centered evaluations of AI-based systems. Specifically, we propose the *sociotechnical*, *ecological*, and *individual* evaluation dimensions, which we describe next.

First, the sociotechnical dimension assesses the suitability of the AI model choice based on the social factors and performance of the model. We align this evaluation dimension with the foundation layer in Suresh et al.'s model design [105]. The foundation layer is proposed to be domain-agnostic and supports a variety of use cases. The sociotechnical evaluation dimension recognizes the social and financial resources required to choose certain AI models, which determines a set of models that could be chosen. A comparison of the AI models within this set—models having a substantive opportunity to be used in the application—on the specific evaluation metrics determines the model choice. The specific metrics to assess the social feasibility could be leveraged from the human resource literature [89] or the AI scholarship jointly optimizing performance with financial cost [45]. The AI model performance evaluation will also be in this dimension based on task-specific [37] or holistic [61] evaluation metrics.

Second, the ecological dimension considers the application domain and assesses the suitability of human infrastructure to mitigate AI fallibility. This aligns with Suresh et al.'s subfloor layer, which recognizes the importance of domain-specificity in determining participant stakeholders in the design of foundational models [105]. Based on domain stakes, this dimension helps a human-centered evaluation guide the human actors whose (capabilities) assessment should be conducted when conducting human-centered evaluations. Here, it may benefit from drawing a correlation between domain stakes and the required actors in the associated infrastructure. We

recommend building on the indices offered by Folbre on evaluating the division of care responsibility across genders [29] to evaluate the wellbeing of different actors involved in the infrastructure. Based on such a metric, we could better understand an AI-based intervention's strengths and limitations, such as compensation and time spent at work. Modeling this relationship between the various actors may help identify and mitigate possible inequitable or arduous work distribution among them.

Third, the individual dimension assesses human capabilities. This dimension is primarily focused on the ends individuals are able to achieve through AI usage. Here, we align with Suresh et al.'s surface dimension, which considers the downstream tasks where the AI model is planned to be used. This dimension recognizes all the humans in the loop and qualitatively evaluates their aspirations and the significant opportunities to achieve them through AI usage. In this dimension, we argue that human-centered evaluations should adopt a broader evaluative focus, which may suffice for individuals looking only to augment their human capital and/or capabilities. To conduct the assessment, we draw from existing capability evaluation methods in literature, like a recent framework formalizing the model of capability assessment through AI usage [64].

Our three evaluation dimensions go hand-in-hand with each other. The aspirations, which are embedded within larger ecology [57], are determined by the domain stakes, which, in turn, determine the AI model choice. A human-centered evaluation conducted across these three dimensions could help better understand the strengths and limitations of an AI-based intervention from a more expansive lens.

7 CONCLUSION

HCI scholarship is considering ways to conduct human-centered evaluations of AI. Our study contributes to these growing conversations. In collaboration with a not-for-profit public health organization with operations in India, we conducted a mixed-methods study and implemented an AI-based system to reduce the workload on the care workers. Leveraging Sen's capability approach as our analytical lens, we found a sociotechnical gap between the care workers' broader aspirations—which they wanted to achieve through the AI-based system—and the relatively narrow ways existing human-centered evaluation metrics defined a system's success. We identified the reasons for this gap and shed light on the role of social factors in determining AI model choice and the high-stakes nature of the application. We argue that the focus of human-centered evaluations should be on assessing AI's success in expanding human capabilities. We end by discussing the three dimensions that HCI researchers and practitioners should consider when conducting human-centered evaluations that help expand the focus to relatively broader achievements that AI could enable among individuals.

ACKNOWLEDGMENTS

We would like to express our deep gratitude to the MSEs who gave us their time and trusted us with their stories. We would also like to thank Preeti Raju and Sneha Vemireddi, who helped build and nurture our relationship with the MSE team. We are deeply grateful to Anubhav Arora, Shirley Yan, Sreeram Ramasubramanian, and team members at The Agency Fund for providing guidance whenever we hit roadblocks and helping ideate our future research goals. A special thanks to our colleagues at the Tandem Lab, Amy Chen, Bill Thies, Jai Moondra, and Mohit Jain, for suggesting potential contributions we could make through this paper. We would like to thank the Design team at Noora Health India Private Limited for providing us with the icons for our figures. This work was funded by the National Science Foundation under the award 2047726. The experimentation credits were provided by OpenAI. Finally, we would like to thank our anonymous reviewers whose valuable feedback helped us understand the strengths of our work and make a focused intellectual contribution to CHI.

REFERENCES

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Mark S Ackerman. 2000. The intellectual challenge of CSCW: the gap between social requirements and technical feasibility. *Human–Computer Interaction* 15, 2-3 (2000), 179–203.

[3] Sudhir Anand and Amartya Sen. 1994. Human Development Index: Methodology and Measurement. (1994).

[4] Maria Antoniak, Aakanksha Naik, Carla S Alvarado, Lucy Lu Wang, and Irene Y Chen. 2024. NLP for Maternal Healthcare: Perspectives and Guiding Principles in the Age of LLMs. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1446–1463.

[5] Ian Arawjo. 2024. *LLM Wrapper Papers are Hurting HCI Research.* https://ianarawjo.medium.com/llm-wrapper-papers-are-hurting-hci-research-8ad416a5d59a

[6] Marianne Aubin Le Quéré, Hope Schroeder, Casey Randazzo, Jie Gao, Ziv Epstein, Simon Tangi Perrault, David Mimno, Louise Barkhuus, and Hanlin Li. 2024. LLMs as Research Tools: Applications and Evaluations in HCI Data Work. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–7.

[7] John W. Ayers, Zechariah Zhu, Adam Poliak, Eric C. Leas, Mark Dredze, Michael Hogarth, and Davey M. Smith. 2023. Evaluating Artificial Intelligence Responses to Public Health Questions. *JAMA Network Open* 6, 6 (June 2023), e2317517. https://doi.org/10.1001/jamanetworkopen.2023.17517

[8] Agathe Balayn, Natasa Rikalo, Jie Yang, and Alessandro Bozzon. 2023. Faulty or Ready? Handling Failures in Deep-Learning Computer Vision Models until Deployment: A Study of Practices, Challenges, and Needs. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.

[9] Luciano Baresi, Matteo Camilli, Tommaso Dolci, and Giovanni Quattrocchi. 2024. A Conceptual Framework for Quality Assurance of LLM-based Socio-critical Systems. In *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering* (Sacramento, CA, USA) *(ASE '24)*. Association for Computing Machinery, New York, NY, USA, 2314–2318. https://doi.org/10.1145/3691620.3695306

[10] Solon Barocas, Anhong Guo, Ece Kamar, Jacquelyn Krones, Meredith Ringel Morris, Jennifer Wortman Vaughan, W Duncan Wadsworth, and Hanna Wallach. 2021. Designing disaggregated evaluations of ai systems: Choices, considerations, and tradeoffs. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 368–378.

[11] 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

[12] Justus Bogner, Roberto Verdecchia, and Ilias Gerostathopoulos. 2021. Characterizing technical debt and antipatterns in AI-based systems: A systematic mapping study. In *2021 IEEE/ACM International Conference on Technical Debt (TechDebt)*. IEEE, 64–73.

[13] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

[14] Ashley Boone, Annabel Rothschild, Xander Koo, Grace Pfohl, Alyssa Sheehan, Betsy DiSalvo, Christopher A Le Dantec, and Carl DiSalvo. 2024. Reimagining Meaningful Data Work through Citizen Science. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (2024), 1–26.

[15] Claus Bossen, Yunan Chen, and Kathleen H. Pine. 2019. The emergence of new data work occupations in healthcare: The case of medical scribes. *International Journal of Medical Informatics* 123 (March 2019), 76–83. https://doi.org/10.1016/j.ijmedinf.2019.01.001

[16] Vikram Kamath Cannanure, Eloísa Ávila-Uribe, Tricia Ngoon, Yves Adji, Sharon Wolf, Kaja Jasińska, Timothy Brown, and Amy Ogan. 2022. "We dream of climbing the ladder; to get there, we have to do our job better": Designing for Teacher Aspirations in rural Côte d'Ivoire. In *Proceedings of the 5th ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies*. 122–138.

[17] Tara Capel and Margot Brereton. 2023. What is human-centered about human-centered AI? A map of the research landscape. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–23.

[18] Stevie Chancellor, Eric PS Baumer, and Munmun De Choudhury. 2019. Who is the" human" in human-centered machine learning: The case of predicting mental health from social media. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–32.

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

[20] Vedant Das Swain, Javier Hernandez, Brian Houck, Koustuv Saha, Jina Suh, Ahad Chaudhry, Tenny Cho, Wendy Guo, Shamsi Iqbal, and Mary P Czerwinski. 2023. Focused time saves nine: Evaluating computer–assisted protected time for hybrid information work. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.

[21] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. 2020. A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[22] David De Cremer and Garry Kasparov. 2021. AI should augment human intelligence, not replace it. *Harvard Business Review* 18, 1 (2021).

[23] Fabrizio Dell'Acqua, Edward McFowland III, Ethan R Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R Lakhani. 2023. Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. *Harvard Business School Technology & Operations Mgt. Unit Working Paper* 24-013 (2023).

[24] Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[25] P Alex Dow, Jennifer Wortman Vaughan, Solon Barocas, Chad Atalla, Alexandra Chouldechova, and Hanna Wallach. 2024. Dimensions of Generative AI Evaluation Design. *arXiv preprint arXiv:2411.12709* (2024).

[26] Emile Durkheim. 2018. The division of labor in society. In *Social stratification*. Routledge, 217–222.

[27] Madeleine Clare Elish and Elizabeth Anne Watkins. [n. d.]. Repairing Innovation: A Study of Integrating AI in Clinical Care. ([n. d.]).

[28] 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.

[29] Nancy Folbre. 2006. Measuring care: Gender, empowerment, and the care economy. *Journal of human development* 7, 2 (2006), 183–199.

[30] Bill Gaver, Tony Dunne, and Elena Pacenti. 1999. Design: cultural probes. *interactions* 6, 1 (1999), 21–29.

[31] R Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. 2020. Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from?. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 325–336.

[32] Jayati Ghosh. 2022. What Do We Really Know about Productivity Differentials across Countries? *Review of Radical Political Economics* 54, 4 (2022), 397–410.

[33] Mary L Gray and Siddharth Suri. 2019. *Ghost work: How to stop Silicon Valley from building a new global underclass*. Eamon Dolan Books.

[34] Venkat Gudivada, Amy Apon, and Junhua Ding. 2017. Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations. *International Journal on Advances in Software* 10, 1 (2017), 1–20.

[35] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. 1998. Support vector machines. *IEEE Intelligent Systems and their applications* 13, 4 (1998), 18–28.

[36] Shanee Honig and Tal Oron-Gilad. 2018. Understanding and resolving failures in human-robot interaction: Literature review and model development. *Frontiers in psychology* 9 (2018), 861.

[37] Mohammad Hossin and Md Nasir Sulaiman. 2015. A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process* 5, 2 (2015), 1.

[38] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276* (2024).

[39] Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren. 2019. *Automated machine learning: methods, systems, challenges*. Springer Nature.

[40] 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*. ACM, Yokohama Japan, 1–21. https://doi.org/10.1145/3411764.3445130

[41] Azra Ismail, Divy Thakkar, Neha Madhiwalla, and Neha Kumar. 2023. Public Health Calls for/with AI: An Ethnographic Perspective. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (Sept. 2023), 1–26. https://doi.org/10.1145/3610203

[42] Adam Tauman Kalai and Santosh S. Vempala. 2024. Calibrated Language Models Must Hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*. ACM, Vancouver BC Canada, 160–171. https://doi.org/10.1145/3618260.3649777

[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*. 1–18.

[44] Sayash Kapoor, Rishi Bommasani, Kevin Klyman, Shayne Longpre, Ashwin Ramaswami, Peter Cihon, Aspen Hopkins, Kevin Bankston, Stella Biderman, Miranda Bogen, et al. 2024. On the Societal Impact of Open Foundation Models. *arXiv preprint arXiv:2403.07918* (2024).

[45] Sayash Kapoor, Benedikt Stroebl, Zachary S Siegel, Nitya Nadgir, and Arvind Narayanan. 2024. Ai agents that matter. *arXiv preprint arXiv:2407.01502* (2024).

[46] Naveena Karusala, Azra Ismail, Karthik S Bhat, Aakash Gautam, Sachin R Pendse, Neha Kumar, Richard Anderson, Madeline Balaam, Shaowen Bardzell, Nicola J Bidwell, et al. 2021. The future of care work: towards a radical politics of care in CSCW research and practice. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*. 338–342.

[47] Naveena Karusala, Shirley Yan, Nupoor Rajkumar, and Richard Anderson. 2023. Speculating with Care: Worker-centered Perspectives on Scale in a Chat-based Health Information Service. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–26.

[48] Magnus Høholt Kaspersen, Line Have Musaeus, Karl-Emil Kjær Bilstrup, Marianne Graves Petersen, Ole Sejer Iversen, Christian Dindler, and Peter Dalsgaard. 2024. From Primary Education to Premium Workforce: Drawing on K-12 Approaches for Developing AI Literacy. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–16.

[49] Aman Khullar, Priyadarshi Hitesh, Shoaib Rahman, Deepak Kumar, Rachit Pandey, Praveen Kumar, Rajeshwari Tripathi, Prince Prince, Ankit Akash Jha, Himanshu Himanshu, and Aaditeshwar Seth. 2021. Costs and Benefits of Conducting Voice-based Surveys Versus Keypress-based Surveys on Interactive Voice Response Systems. In *ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS)*. ACM, Virtual Event Australia, 288–298. https://doi.org/10.1145/3460112.3471963

[50] Aman Khullar, Paramita Panjal, Rachit Pandey, Abhishek Burnwal, Prashit Raj, Ankit Akash Jha, Priyadarshi Hitesh, R Jayanth Reddy, Himanshu Himanshu, and Aaditeshwar Seth. 2021. Experiences with the Introduction of AI-based Tools for Moderation Automation of Voice-based Participatory Media Forum. In *India HCI 2021*. ACM, Virtual Event India, 30–39. https://doi.org/10.1145/3506469.3506473

[51] Aman Khullar, M Santosh, Praveen Kumar, Shoaib Rahman, Rajeshwari Tripathi, Deepak Kumar, Sangeeta Saini, Rachit Pandey, and Aaditeshwar Seth. 2021. Early Results from Automating Voice-based Question-Answering Services Among Low-income Populations in India. In *ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS)*. ACM, Virtual Event Australia, 79–87. https://doi.org/10.1145/3460112.3471946

[52] Jini Kim and Hajun Kim. 2024. Unlocking Creator-AI Synergy: Challenges, Requirements, and Design Opportunities in AI-Powered Short-Form Video Production. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–23.

[53] Lauren Klein and Catherine D'Ignazio. 2024. Data Feminism for AI. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 100–112.

[54] Thomas Kosch and Sebastian Feger. 2024. Risk or Chance? Large Language Models and Reproducibility in HCI Research. *Interactions* 31, 6 (2024), 44–49.

[55] 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.

[56] Neha Kumar, Naveena Karusala, Azra Ismail, and Anupriya Tuli. 2020. Taking the long, holistic, and intersectional view to women's wellbeing. *ACM Transactions on Computer-Human Interaction (TOCHI)* 27, 4 (2020), 1–32.

[57] Neha Kumar, Marisol Wong-Villacres, Naveena Karusala, Aditya Vishwanath, Arkadeep Kumar, and Azra Ismail. 2019. Aspirations-based design. In *Proceedings of the tenth international conference on information and communication technologies and development*. 1–11.

[58] J Steven Landefeld, Eugene P Seskin, and Barbara M Fraumeni. 2008. Taking the pulse of the economy: Measuring GDP. *Journal of Economic Perspectives* 22, 2 (2008), 193–216.

[59] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. The parable of Google Flu: traps in big data analysis. *science* 343, 6176 (2014), 1203–1205.

[60] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.

[61] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic Evaluation of Language Models. *Transactions on Machine Learning Research* (2022).

[62] Q Vera Liao and Ziang Xiao. 2023. Rethinking model evaluation as narrowing the socio-technical gap. *arXiv preprint arXiv:2306.03100* (2023).

[63] Houjiang Liu, Anubrata Das, Alexander Boltz, Didi Zhou, Daisy Pinaroc, Matthew Lease, and Min Kyung Lee. 2024. Human-centered NLP Fact-checking: Co-Designing with Fact-checkers using Matchmaking for AI. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (2024), 1–44.

[64] Alex John London and Hoda Heidari. 2024. Beneficent intelligence: a capability approach to modeling benefit, assistance, and associated moral failures through AI systems. *Minds and Machines* 34, 4 (2024), 41.

[65] Sasha Luccioni, Boris Gamazaychikov, Sara Hooker, Régis Pierrard, Emma Strubell, Yacine Jernite, and Carole-Jean Wu. 2024. Light bulbs have energy ratings—so why can't AI chatbots? *Nature* 632, 8026 (2024), 736–738.

[66] Zilin Ma, Yiyang Mei, Yinru Long, Zhaoyuan Su, and Krzysztof Z Gajos. 2024. Evaluating the Experience of LGBTQ+ People Using Large Language Model Based Chatbots for Mental Health Support. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–15.

[67] Alva Markelius, Connor Wright, Joahna Kuiper, Natalie Delille, and Yu-Ting Kuo. 2024. The mechanisms of AI hype and its planetary and social costs. *AI and Ethics* (2024), 1–16.

[68] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2022. A Survey on Bias and Fairness in Machine Learning. *Comput. Surveys* 54, 6 (July 2022), 1–35. https://doi.org/10.1145/3457607

[69] Carlo Mervich. 2020. *The human infrastructure of artificial intelligence*. Master's thesis. University of Twente.

[70] Michael Muller, Ingrid Lange, Dakuo Wang, David Piorkowski, Jason Tsay, Q. Vera Liao, Casey Dugan, and Thomas Erickson. 2019. How Data Science Workers Work with Data: Discovery, Capture, Curation, Design, Creation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–15. https://doi.org/10.1145/3290605.3300356

[71] Naja Holten Møller, Claus Bossen, Kathleen H. Pine, Trine Rask Nielsen, and Gina Neff. 2020. Who does the work of data? *Interactions* 27, 3 (April 2020), 52–55. https://doi.org/10.1145/3386389

[72] Gireen Naidu, Tranos Zuva, and Elias Mmbongeni Sibanda. 2023. A review of evaluation metrics in machine learning algorithms. In *Computer Science On-line Conference*. Springer, 15–25.

[73] 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*. ACM, Yokohama Japan, 1–20. https://doi.org/10.1145/3411764.3445420

[74] Ilse Oosterlaken. 2012. The capability approach, technology and design: Taking stock and looking ahead. In *The capability approach, technology and design*. Springer, 3–26.

[75] OpenAI. 2024. *Usage policies*. https://openai.com/policies/usage-policies/

[76] International Labor Organization. 2024. *Decent work and the care economy*. https://www.ilo.org/resource/conference-paper/decent-work-and-care-economy

[77] Siddiqur R Osmani. 2016. The capability approach and human development: some reflections. *United Nations Development Porgramme: New York, NY, USA* (2016).

[78] 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 *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–13.

[79] Samir Passi and Phoebe Sengers. 2020. Making data science systems work. *Big Data & Society* 7, 2 (July 2020), 205395172093960. https://doi.org/10.1177/2053951720939605

[80] Kathleen Pine, Claus Bossen, Naja Holten Møller, Milagros Miceli, Alex Jiahong Lu, Yunan Chen, Leah Horgan, Zhaoyuan Su, Gina Neff, and Melissa Mazmanian. 2022. Investigating Data Work Across Domains: New Perspectives on the Work of Creating Data. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. ACM, New Orleans LA USA, 1–6. https://doi.org/10.1145/3491101.3503724

[81] Kathleen H Pine and Claus Bossen. 2020. Good organizational reasons for better medical records: The data work of clinical documentation integrity specialists. *Big Data & Society* 7, 2 (July 2020), 205395172096561. https://doi.org/10.1177/2053951720965616

[82] Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, Melbourne, Australia, 784–789. https://doi.org/10.18653/v1/P18-2124

[83] Pragnya Ramjee, Mehak Chhokar, Bhuvan Sachdeva, Mahendra Meena, Hamid Abdullah, Aditya Vashistha, Ruchit Nagar, and Mohit Jain. 2024. ASHABot: An LLM-Powered Chatbot to Support the Informational Needs of Community Health Workers. *arXiv preprint arXiv:2409.10913* (2024).

[84] Rasmus V Rasmussen and Michael A Trick. 2008. Round robin scheduling–a survey. *European Journal of Operational Research* 188, 3 (2008), 617–636.

[85] Shahra Razavi. 2007. The political and social economy of care in a development context: Conceptual issues, research questions and policy options. *Trabajo y empleo* (2007).

[86] Sarah T Roberts. 2020. Behind the Screen: Content Moderation in the Shadows of Social Media.

[87] Ingrid Robeyns and Morten Fibieger Byskov. 2023. The Capability Approach. In *The Stanford Encyclopedia of Philosophy* (Summer 2023 ed.), Edward N. Zalta and Uri Nodelman (Eds.). Metaphysics Research Lab, Stanford University.

[88] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115 (2015), 211–252.

[89] Shefali Sachdeva. 2015. HR metrics for value of human resources: gaps and direction. *International Journal of Research in Social Sciences* 5, 2 (2015), 496–508.

[90] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M 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*. ACM, Yokohama Japan, 1–15. https://doi.org/10.1145/3411764.3445518

[91] Nithya Sambasivan and Rajesh Veeraraghavan. 2022. The deskilling of domain expertise in AI development. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–14.

[92] Elizabeth Seger, Aviv Ovadya, Divya Siddarth, Ben Garfinkel, and Allan Dafoe. 2023. Democratising AI: Multiple meanings, goals, and methods. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 715–722.

[93] Joshua Paolo Seguin, Delvin Varghese, Misita Anwar, Tom Bartindale, and Patrick Olivier. 2022. Co-designing digital platforms for volunteer-led migrant community welfare support. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference*. 247–262.

[94] Amartya Sen. 1999. Development as Freedom: Oxford univerisity Press.

[95] Amartya Kumar Sen. 1997. Human capital and human capability. *World development* 25, 12 (1997), 1959–1961.

[96] Aaditeshwar Seth. 2022. Overcoming Paradigms That Disempower. In *Technology and (Dis) Empowerment: A Call to Technologists*. Emerald Publishing Limited, 145–159.

[97] Vishal Sharma, Neha Kumar, and Bonnie Nardi. 2023. Post-growth Human–Computer Interaction. *ACM Transactions on Computer-Human Interaction* 31, 1 (2023), 1–37.

[98] Chenxi Shi, Penghao Liang, Yichao Wu, Tong Zhan, and Zhengyu Jin. 2024. Maximizing User Experience with LLMOps-Driven Personalized Recommendation Systems. *arXiv preprint arXiv:2404.00903* (2024).

[99] Christos Skevas, Nicolás Pérez de Olaguer, Albert Lleó, David Thiwa, Ulrike Schroeter, Inês Valente Lopes, Luca Mautone, Stephan J Linke, Martin Stephan Spitzer, Daniel Yap, et al. 2024. Implementing and evaluating a fully functional AI-enabled model for chronic eye disease screening in a real clinical environment. *BMC ophthalmology* 24, 1 (2024), 51.

[100] Irene Solaiman, Zeerak Talat, William Agnew, Lama Ahmad, Dylan Baker, Su Lin Blodgett, Canyu Chen, Hal Daumé III, Jesse Dodge, Isabella Duan, et al. 2023. Evaluating the social impact of generative ai systems in systems and society. *arXiv preprint arXiv:2306.05949* (2023).

[101] Ian René Solano-Kamaiko, Dibyendu Mishra, Nicola Dell, and Aditya Vashistha. 2024. Explorable Explainable AI: Improving AI Understanding for Community Health Workers in India. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–21. https://doi.org/10.1145/3613904.3642733

[102] Elizabeth A Stanton. 2007. Engendering human development: A critique of the UNDP's Gender-related Development Index. (2007).

[103] Yuling Sun, Xiaojuan Ma, Silvia Lindtner, and Liang He. 2023. Data Work of Frontline Care Workers: Practices, Problems, and Opportunities in the Context of Data-Driven Long-Term Care. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (April 2023), 1–28. https://doi.org/10.1145/3579475

[104] Harini Suresh and John Guttag. 2021. A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–9.

[105] Harini Suresh, Emily Tseng, Meg Young, Mary Gray, Emma Pierson, and Karen Levy. 2024. Participation in the age of foundation models. In *The 2024 ACM*

Conference on Fairness, Accountability, and Transparency. 1609–1621.

[106] 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*. ACM, New Orleans LA USA, 1–16. https://doi.org/10.1145/3491102.3501868

[107] Paola Tubaro, Antonio A Casilli, and Marion Coville. 2020. The trainer, the verifier, the imitator: Three ways in which human platform workers support artificial intelligence. *Big Data & Society* 7, 1 (Jan. 2020), 205395172091977. https://doi.org/10.1177/2053951720919776

[108] Mahbub Ul Haq. 2003. The birth of the human development index. *Readings in human development* 2 (2003), 127–137.

[109] Hanna Wallach, Meera Desai, Nicholas Pangakis, A Feder Cooper, Angelina Wang, Solon Barocas, Alexandra Chouldechova, Chad Atalla, Su Lin Blodgett, Emily Corvi, et al. 2024. Evaluating Generative AI Systems is a Social Science Measurement Challenge. *arXiv preprint arXiv:2411.10939* (2024).

[110] Dakuo Wang, Liuping Wang, Zhan Zhang, Ding Wang, Haiyi Zhu, Yvonne Gao, Xiangmin Fan, and Feng Tian. 2021. "Brilliant AI Doctor" in Rural China: Tensions and Challenges in AI-Powered CDSS Deployment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–18. https://doi.org/10.1145/3411764.3445432 arXiv:2101.01524 [cs].

[111] Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024. Human-LLM collaborative annotation through effective verification of LLM labels. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–21.

[112] Robert Wolfe and Tanushree Mitra. 2024. The Impact and Opportunities of Generative AI in Fact-Checking. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 1531–1543. https://doi.org/10.1145/3630106.3658987

[113] Yunpeng Xiao, Kyrie Zhixuan Zhou, Yueqing Liang, and Kai Shu. 2024. Understanding the concerns and choices of public when using large language models for healthcare. http://arxiv.org/abs/2401.09090 arXiv:2401.09090 [cs].

[114] Ziang Xiao, Wesley Hanwen Deng, Michelle S Lam, Motahhare Eslami, Juho Kim, Mina Lee, and Q Vera Liao. 2024. Human-Centered Evaluation and Auditing of Language Models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–6.

[115] Ziang Xiao, Q Vera Liao, Michelle Zhou, Tyrone Grandison, and Yunyao Li. 2023. Powering an ai chatbot with expert sourcing to support credible health information access. In *Proceedings of the 28th international conference on intelligent user interfaces*. 2–18.

[116] Deepika Yadav, Prerna Malik, Kirti Dabas, and Pushpendra Singh. 2021. Illustrating the Gaps and Needs in the Training Support of Community Health Workers in India. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[117] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems* 36 (2023), 46595–46623.