Title

NLP for Maternal Healthcare: Perspectives and Guiding Principles in the Age of LLMs

Authors

Antoniak, Maria

Naik, Aakanksha

Alvarado, Carla S

et al.

Publication Date

2024-06-03

DOI

10.1145/3630106.3658982

NLP for Maternal Healthcare: Perspectives and Guiding Principles in the Age of LLMs

MARIA ANTONIAK, Allen Institute for AI, USA
AAKANKSHA NAIK, Allen Institute for AI, USA
CARLA S. ALVARADO, Association of American Medical Colleges, Center for Health Justice, USA
LUCY LU WANG, University of Washington, Allen Institute for AI, USA
IRENE CHEN, University of California, Berkeley and University of California, San Francisco, USA

Ethical frameworks for the use of natural language processing (NLP) are urgently needed to shape how large language models (LLMs) and similar tools are used for healthcare applications. Healthcare faces existing challenges including the balance of power in clinician-patient relationships, systemic health disparities, historical injustices, and economic constraints. Drawing directly from the voices of those most affected, and focusing on a case study of a specific healthcare setting, we propose a set of guiding principles for the use of NLP in maternal healthcare. We led an interactive session centered on an LLM-based chatbot demonstration during a full-day workshop with 39 participants, and additionally surveyed 30 healthcare workers and 30 birthing people about their values, needs, and perceptions of NLP tools in the context of maternal health. We conducted quantitative and qualitative analyses of the survey results and interactive discussions to consolidate our findings into a set of guiding principles. We propose nine principles for ethical use of NLP for maternal healthcare, grouped into three themes: (i) recognizing contextual significance (ii) holistic measurements, and (iii) who/what is valued. For each principle, we describe its underlying rationale and provide practical advice. This set of principles can provide a methodological pattern for other researchers and serve as a resource to practitioners working on maternal health and other healthcare fields to emphasize the importance of technical nuance, historical context, and inclusive design when developing NLP technologies for clinical use.

CCS Concepts: • **Applied computing → Health informatics**; • **Computing methodologies → Natural language processing**; • **Social and professional topics → Medical technologies**.

Additional Key Words and Phrases: maternal health, natural language processing, large language models, ethical guidelines

As natural language processing (NLP) methods and large language models (LLMs) have increased in size and performance, so has hype and excitement increased about their clinical use [89]. Recent work experimenting with generative LLMs has found promising results in comparison to physicians, including for tasks such as addressing public health concerns [5], answering care-seekers' questions [4], and engaging in diagnostic dialogues [90]. But healthcare and ethics researchers have also highlighted the safety risks [56], biases [102], inaccuracies [13], and other problems arising from NLP tools applied to healthcare contexts, as well as general risks of LLMs [9, 99]. Given the sensitivity of medical data and the potential for harmful negative outcomes, deciding when and how to employ NLP technologies in clinical settings requires careful consideration.

Two key challenges complicate these decisions. First, numerous stakeholders have a shared interest in the development and use of these technologies; care-seekers, clinicians, researchers, hospital administration, and other groups can all

Authors' addresses: Maria Antoniak, Allen Institute for AI, USA; Aakanksha Naik, Allen Institute for AI, USA; Carla S. Alvarado, Association of American Medical Colleges, Center for Health Justice, USA; Lucy Lu Wang, University of Washington, Allen Institute for AI, USA; Irene Chen, University of California, Berkeley and University of California, San Francisco, USA.

provide design and decision-making input, but these voices are not equally represented, may conflict, and are rarely brought together during decisions around system design and implementation. Second, most research developing ethical guidelines has studied high level applications of machine learning for healthcare, across (i) a range of medical topics and (ii) across a range of machine learning methods. There is a gap for research grounded in a specific healthcare setting where risks, benefits, and stakeholder perceptions can be fully explored, focusing specifically on NLP technologies and the new risks posed by LLMs.

In this work, we focus on a case study of *maternal healthcare* in the United States. We employ a participatory design framework [58] to amplify the voices of diverse stakeholder groups who have previously been underrepresented in the discussion around use of NLP technologies in healthcare. Drawing on the input from these groups, we identify a set of *guiding principles* that support well-grounded and ethical uses of NLP and LLM technologies in maternal healthcare.

We choose maternal health for three reasons. First, there are many prior research studies and applications of NLP methods focused on maternal healthcare [36, 46, 49, 94]. Second, pregnancy and childbirth are common events that often comprise a person's sole or major interaction with the healthcare system, increasing the significance and also abundance of perspectives on this topic. Third, maternal health is a "perfect storm" of healthcare vulnerabilities, with historical biases and power dynamics influencing care; for example, maternal health in the U.S. has received much attention in recent years due to the high morbidity and mortality of birthing people and significant racial inequities [25, 53]. Focusing on maternal health allows us to map our investigation of NLP applications over a set of *specific* and *grounded* opportunities and risks.

We solicited key considerations about an LLM-based chatbot and NLP tools from diverse stakeholders, including clinicians, birthing people, researchers, community health workers, government and non-profit workers, and others. We collected perspectives in two main ways (Figure 1). First, we introduced stakeholders to NLP technologies, allowing them to surface concerns, opportunities, and discussion through interaction with these technologies in a controlled setting. Second, we conducted post-interaction surveys to gather information about priorities, opportunities, and risks.

Through analysis of the collected data, iterative rounds of discussion, and literature review, we consolidated and identified nine guiding principles for the use of NLP in maternal healthcare, organized according to three themes: *context*, *measurements*, and *values*. These themes are inspired by our maternal health case study but provide a transferable blueprint for mapping tensions in other healthcare areas, supporting NLP practitioners and healthcare administrators on decisions related to the design and deployment of NLP tools.

1 RELATED WORK

1.1 NLP and LLMs for Healthcare

NLP methods have been applied to many kinds of health data, both in research and in industry applications. Traditionally, NLP for healthcare has focused on structured tasks like information extraction from clinical notes [76, 81, 97] and outcome prediction [59, 80, 95], or on public health surveillance of social media [6, 34, 61]. Technical advances in LLMs have shifted this focus, with recent work applying generative NLP tools to a wide variety of question-answering use cases, including engaging in diagnostic dialogues [90], addressing medical scenarios [63, 92], and generating responses for patients asking questions on social media sites [4]. Research has also focused on developing large datasets [31] or studying clinician adoption of LLMs [45, 82]. On the applied side, as one example, Epic Systems, the leading provider of electronic health record (EHR) systems in the United States, is bringing GPT-4 into EHRs to help clinicians communicate with patients [91]. And in practice, everyday users are already exploring their healthcare questions with generic

LLM-based chat tools; according to one analysis of a publicly-released dataset of user-GPT conversation histories, approximately 3% of user queries are health-related [66].

These use cases have fueled both optimism [50, 93] and scrutiny [93, 101]. Within the fairness, accountability, and transparency (FAccT) community, researchers have studied many general risks and challenges arising with LLMs [9, 99] and also highlighted risks specific to healthcare settings. For example, recent work on ethical uses of machine learning for healthcare has studied bias measurement [102], explainability methods [7, 67, 69], dataset documentation gaps [72], and procedures for ethical implementations and deployments [78]. Compared to research on other application areas, these works emphasize the particular privacy issues associated with clinical data [102], the importance of optimizing for the right outcome variables (health outcomes rather than institutional costs) [64], and interactions between healthcare workers and AI systems [32] that require explainability [7, 67, 69].

Most of these works addressing healthcare contexts study machine learning methods more broadly, rather than focusing on NLP methods and the specific challenges arising from new tools like LLMs. We build on these works by focusing on NLP tools for maternal health and by eliciting perceptions from multiple stakeholder groups.

1.2 Ethical Guidelines for NLP/AI and Healthcare

Prior work has developed important guidelines and frameworks for the use of machine learning (ML) and artificial intelligence methods for healthcare. These works take a general view of healthcare and are often focused on ML practitioners and their processes. For example, Mccradden et al. [54] focuses on ethical guidelines for ML-informed clinical decision-making, and Chen et al. [22] and Wiens et al. [100] provide overviews of the ML/NLP development pipeline and recommendations focused on each pipeline step. The recommendations in these works are based on literature reviews and broad sets of healthcare examples. While most of these guidelines consider healthcare broadly, Sendak et al. [78] design guidelines based on the deployment of a specific sepsis-detection machine learning tool, and Petti et al. [68] focused on developing ethical guidelines for the use of NLP and AI methods for early detection of Alzheimer's disease. Our work builds most directly on the latter study, as we also focus on linguistic data and NLP methods.

We build on these works by combining in a novel way the following goals: (i) focusing on NLP and new LLM technologies, (ii) grounding our work in maternal health as a specific healthcare context, and (iii) soliciting direct feedback from multiple stakeholder groups. In particular, we elicit values from various stakeholders using a framework introduced by Jakesch et al. [40], which aims to model the "human process" at the center of healthcare [33]. As recommended by Rajkomar et al. [71] and following prior work [96], we follow a research process that draws on *participatory design*, a collaborative design approach that involves stakeholders and end users in the design process for new technologies [58], such that their needs drive design decisions and result in tools that better address their concerns and challenges [28, 39, 48, 85]. We engage with these stakeholder groups not to "solve" issues with NLP tools but to unearth concerns and themes [83]. We also make use of participatory design in the formation of our guiding principles; uniquely, we survey stakeholders connected to a specific healthcare topic to better illuminate tensions and priorities for NLP usage for healthcare.

1.3 Maternal Health: Vulnerabilities and NLP Applications

The U.S. is experiencing an urgent and worsening maternal mortality crisis. Rates of pregnancy-related deaths and complications have increased over the last thirty years in the U.S., with particularly high rates among Hispanic and Black birthing people and non-U.S. citizens [11, 24, 25, 52, 53]. For example, Black women are three times more likely to die in

childbirth than white women [25, 53]. Most frustratingly, 46% of maternal deaths of Black women and 33% of maternal deaths of white women are estimated as preventable [12], but measurements of these problems are challenging [27]. Additional dangers face birthing people, such as postpartum depression [88], and urgently need supportive solutions.

Researchers and healthcare practitioners have sometimes turned to NLP methods to try to address various challenges in the maternal health space. For example, prior research has used NLP methods to study the prevalence of adverse outcomes via social media data [49], extract lactation information from drug labels [36], predict high-risk pregnancies from electronic health records (EHR) [46], examine how news media discuss pregnancy and exposure to isotretinoin [57], and assist in analysis of phone interviews about breastfeeding [94]. In our review of this prior work in NLP, we observe the following patterns: (i) many studies attempt to surveil and predict psychological states of the birthing person, e.g., predicting postpartum depression, and (ii) most studies use either social media or EHR data, with a bias in studies published at NLP venues towards social media data produced by birthing people.

2 METHODS

In Figure 1, we provide an overview of our methods. Participants first engage with a structured interface to query an LLM-based chatbot. After participants have interacted with this demonstration, they move to discussions (either virtual breakout discussion sessions led by volunteer moderators or independent written comments), and a survey, in which we asked participants to express their perceptions of NLP tools for maternal healthcare. Our study was approved by the institutional review board (IRB) at the Allen Institute for AI.

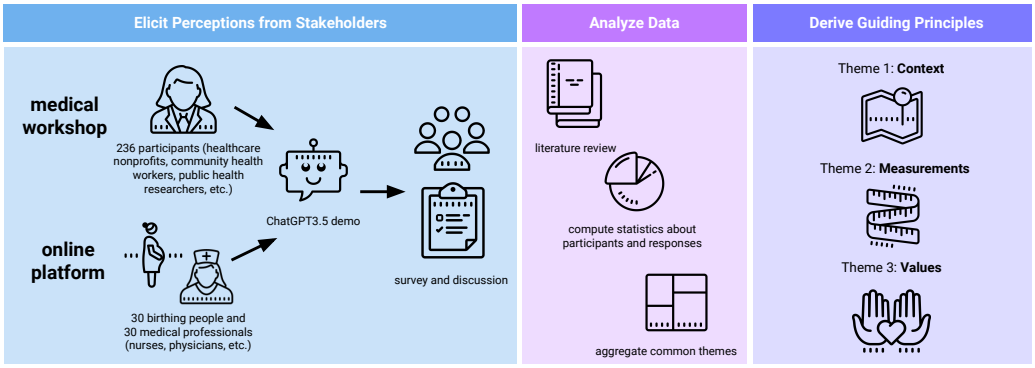

Fig. 1. Study overview, including the three participant cohorts, the chatbot demonstration, and the surveys and discussions followed by analysis and design of the guiding principles.

2.1 Participant Cohorts

We ran this study with three participant cohorts. For full demographic descriptions of these cohorts, see Table 1 in §3.1.

2.1.1 Healthcare workers and birthing people. Using the Prolific[1] online survey platform, we recruited one cohort of birthing people (gave birth in the last five years) and one cohort of healthcare workers (have ever worked in any healthcare profession). We required that all participants be over 18 years old and be located in the U.S. Prolific workers were paid an average of $15/hour.

[1] https://www.prolific.com/

2.1.2 Workshop participants. For the third cohort, we led a live, virtual session with a group of participants who first received training about LLMs and their risks. Our session was held at the end of a virtual workshop hosted by a U.S.-based medical nonprofit. The workshop was themed around NLP and maternal health equity, and in sessions preceding ours, participants were introduced to the basics of NLP, heard research talks about applications of NLP to maternal health, and learned about biases and ethical challenges in NLP tools. The workshop was open to the public, and participants were solicited through the nonprofit's email listserv, social media, and word of mouth. Most but not all of these participants also worked in healthcare, though compared to our cohort from Prolific, these participants more often worked in community and research roles rather than as clinicians (see Table 1).

2.2 Design of chatbot demonstration

We built an interface wrapper (Figure 10) around GPT-3.5 from OpenAI.[2] Our goal with this demonstration was to give participants a guided but unconstrained interaction with a chatbot, gathering their queries, structured feedback, and perceptions related to the chatbot and maternal health. For many of our participants (see Figure 3 in the Appendix), this was their first direct interaction with a modern LLM. We included prominent warnings about (i) privacy and data collection and (ii) the importance of always seeking the advice of a healthcare professional (see Figure 10 in Appendix B). We also provided examples of the kinds of maternal health questions one might ask a chatbot.

2.3 Design of discussion sessions

2.3.1 Workshop participants. Workshop participants were divided into eight virtual breakout rooms and participated in a 30-minute free-form discussion session, moderated by the authors and trained volunteers. Discussion sessions were recorded after obtaining consent from all participants. During the session, participants were shown the following discussion prompts:

(1) How was your experience with the chatbot? What stood out to you about the responses?
(2) What are your dream NLP tools for maternal health? What tools should never be built?
(3) Which maternal health stakeholders (birthing people, nurses, doulas, etc.) would benefit or be hurt by NLP tools?
(4) What principles should guide the use of NLP for maternal health? What should be the goals and guardrails?

We collected video and audio recordings for four out of eight sessions as well as written comments (collected via Mentimeter[3]) from five sessions, which we included in our analysis.

2.3.2 Healthcare workers and birthing people. Like the workshop participants, non-workshop participants first interacted with our chatbot demonstration and then wrote answers to our discussion questions as part of a followup survey. Unlike the workshop participants, non-workshop participants did not participate in discussions with groups or with a moderator but instead completed the study independently.

2.4 Survey design

Our goals in the survey were to (i) elicit participants' general perceptions of NLP tools, (ii) learn about participants' information seeking goals, and (iii) build collaborative rankings of values that should guide NLP uses for maternal health. Surveys given to each cohort were identical except that for birthing people, questions about healthcare careers

[2]Compared to GPT-4, which was also accessible at the time of our study, GPT-3.5 is less expensive and has lower rate limits, which was especially important for the workshop setting, where 100+ people were interacting with the bot at any moment. Our goal was to demonstrate high performance for the bot rather than create a reproducible system; otherwise, an open model would have been preferable.
[3]https://www.mentimeter.com/

were replaced with questions about whether the participant would like their healthcare team to use and/or disclose their use of AI tools. See Appendix C for the full set of survey questions.

Participants were first asked about their generalized trust then asked about their trust in healthcare providers; the format of these questions is drawn from Baughan et al. [8]. Participants were then asked about their familiarity with NLP tools like ChatGPT, as well as their general perceptions of AI; these questions were taken from a frequently reused set by [60]. Next, participants were asked to select five out of ten *ethical values* that should guide the use of NLP for maternal health. The listed values were taken from Jakesch et al. [40] in a study of variations in attitudes towards AI by different demographic groups. We provided definitions (also drawn from Jakesch et al. [40]) and example healthcare applications of NLP systems. Participants were asked about their information seeking needs (where and to whom they turn with maternal health questions) and asked their opinion about the effects of NLP tools on maternal health team members and on their own job (if applicable). Finally, participants answered a series of demographic questions, including specifying any professional role they had ever taken in maternal health (e.g., worked as a midwife).

3 RESULTS

3.1 Survey Results

We release our survey results publicly to support future research.[4] Figure 2 shows how frequently each value was selected by each cohort when asked to "select any five values ... that you think are most important for NLP systems for maternal health." Overall, *safety*, *privacy*, and *performance* were selected more often by birthing people and healthcare workers in comparison to workshop participants, who were more likely to select *inclusiveness*. The birthing people and workshop participants were united in being more likely to select *human autonomy* than the healthcare workers, and less likely than the healthcare workers to select *transparency* and *accountability*. Overall, we do not find large differences in the value selections within cohorts across stratifications for generalized trust, trust in healthcare workers, or generalized AI perceptions (see Appendix A).

Table 1 provides a summary of participant demographics and background. The workshop participants tended to represent the non-profit, community health, research, and governmental sectors in contrast to clinicians such as physicians and nurses that were more common in the Prolific cohort of healthcare workers. The workshop participants represented a more even distribution across racial/ethnic groups, but all cohorts had little representation in the East Asian, South Asian, and Southeast Asian groups.[5] Overall, workshop participants reported less experience with AI and NLP than birthing people and healthcare workers recruited from Prolific (see Appendix A).

3.2 Thematic Analysis of Discussion Sessions

We identified themes and concepts frequently brought up by study participants in workshop breakout discussions, written comments, free-text survey answers, and chatbot queries. We first performed open coding on responses from each source to identify relevant concepts [20], then axial coding to group these concepts under broad themes [23]. The authors collectively discussed these themes and iterated several times before synthesizing them into the final set of three themes and nine guiding principles presented in §4. We summarize these principles in the following section and

[4]We remove demographic features for the workshop participants, as these participants interacted with one another and could potentially identify others' survey responses. Participants consented to the sharing of their responses. All survey data is available in supplementary materials and at a public Github repository.

[5]We attempted to correct this balance on Prolific by supplementing our results with an additional cohort of East Asian, South Asian, and Southeast Asian participants, but we found that the Prolific survey platform included very few such participants who also met our other criteria (e.g., gave birth in the last five years).

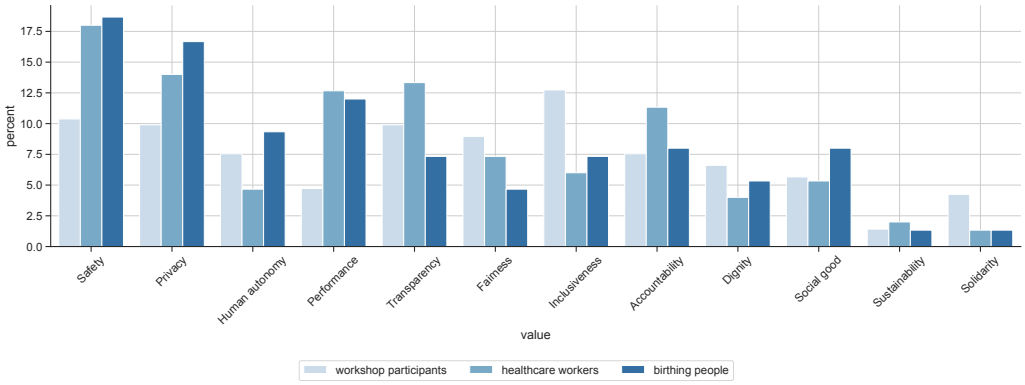

Fig. 2. Frequencies of selected values. Each participant was asked to select five values from a list of 12 curated by Jakesch et al. [40]. Participants were given definitions of the values, also drawn from Jakesch et al. [40]. Birthing people and healthcare workers overall responded more similarly to each other than to the workshop participants, but this was not always the case (e.g., *transparency*, *human autonomy*).

include anonymized quotes from our study participants to highlight stakeholder pain points and further support the inclusion of each principle. Additionally, we provide actionable recommendations to NLP researchers and practitioners based on our own experience and expertise as NLP researchers and healthcare workers. Our thematic analyses also raised several additional points, which we discuss in §5.

4 GUIDING PRINCIPLES

Drawing from the discussion themes, our survey results, and related work (both applied work and high-level ethics guidelines like those by Chen et al. [22] and D'Ignazio and Klein [29]), we develop guiding principles for the responsible use of NLP in maternal health.

We present these principles grouped under three themes: *context*, *measurements*, and *values*. *Context* principles ask practitioners to incorporate the fundamentals and history of maternal health in their applications, *measurement* principles discuss what to optimize for and how to evaluate, and *value* principles address how practitioners should situate user voices and data relevant to their systems.

We summarize each principle below and highlight in participants' own words their importance. In quotes, workshop participants are attributed as W1-10, birthing people as B1-11, and healthcare workers as H1-15.

4.1 Theme 1: Context

Be aware of power dynamics in the care team. D'Ignazio and Klein [29] have argued that *power* should be a central concern of feminist data science, and maternal healthcare has a long and fraught history of shifting power dynamics in the care team [10]. These historical shifts, e.g., male OB/GYNs replacing midwives in the early 20th century [15, 51], have led to the marginalization of some maternal health workers and both good and bad consequences for birthing people. NLP practitioners should know this history and work to improve rather than exacerbate existing hierarchies, being mindful of the impact of tools and automation on midwives, doulas, and other care workers whose placement is already precarious. For example, one of the healthcare workers in our study mentioned that *"an AI chatbot can eliminate the need for a Doula"* (H1), highlighting the vulnerabilities of certain workers. One of the birthing people framed this as

Cohort	Race/Ethnicity	Age	Highest Education	Gender
		Workshop Participants (N = 39)		
38% community nonprofits 27% pop./public health research 24% comm. health/promotara 24% local/state public health 19% healthcare management/admin 16% healthcare services researcher 11% other perinatal healthcare provider 11% other non-healthcare perinatal support 8% doula 8% non-perinatal healthcare provider 13.5% *all other groups*	41% African-American/Black 41% White 16% Hispanic/Latino/a/x 5% South Asian 19% *all other groups*	35% 35-44 30% 25-34 19% 45-54 11% 55-64 5% 65-74	38% MS, MPH, etc. 30% PhD 24% BA, BS, etc. 11% *all other groups*	92% women 5% men 3% no answer 0% non-binary
		Healthcare Workers (N = 30)		
20% nurse 17% pharmacy 10% physician 10% medical tech 10% medical assistant/aide 10% research 23% *all other groups* 33% have worked in maternal/perinatal healthcare	57% White 23% African-American/Black 7% East Asian 7% Southeast Asian 9% *all other groups*	33% 35-44 30% 25-34 10% 18-24 3% 65-74 3% 55-64	50% BA, BS, etc. 17% MS, MPH, etc. 17% Trade School 10% MD, DO, etc. 7% Community College 6% *all other groups*	79% women 21% men 0% non-binary
		Birthing People (N = 30)		
20% have worked in healthcare 7% have worked in maternal/perinatal healthcare	73% White 20% Hispanic/Latino/a/x 17% African-American/Black 12% *all other groups*	53% 25-34 37% 35-44 10% 65-74	33% BA, BS, etc. 30% High school or GED 13% Community College 10% MS, MPH, etc. 10% Trade School 7% PhD 3% Prof. Degree	97% women 7% non-binary 0% men

Table 1. Demographic description of participant cohorts. Healthcare workers and birthing people were recruited from the Prolific platform while the workshop participants took part in multiple training and educational sessions to learn about LLMs, their applications to maternal health, and their risks. Work category, race/ethnicity, highest education, and gender all allowed multiple selections, so the percents for these categories might not sum to 100%.

a harm rather than an opportunity: *"If mothers relied too heavily on AI instead of seeking professional help then the nurses and doulas may see fewer people seeking care"* (B11).

Recommended actions for practitioners:

- Understand historical power shifts in the care team. Consider carefully who you expect to use your tool and who will be impacted (or potentially replaced) by your tool. Rather than designing a chatbot as a replacement for a core member of the care team, consider designing supportive tools that can improve the care team's collaboration.
- Consider these questions: Whose language and data are represented in the training data of the tool? Who will receive predictions and advice? Who will have access to and control? Are you concentrating more power within a single care team role?

Know the politics and implications of your measurements. Using NLP tools for data collection can have unintended consequences that can impact the *safety* and *agency* of birthing people. For example, (i) overemphasis on data collection can draw resources away from building solutions [29], (ii) reproductive health is politicized, and measurements may be used as evidence in support of unexpected agendas, and (iii) focusing only on measuring problems can contribute to "deficit narratives" that blame communities for their own challenges [29]. In our literature review (§1), we found that much prior NLP research for maternal health has focused on data collection and surveillance of birthing people, and participants pushed back against this pattern: *"It's key that the AI is actually able to provide solutions to problems that can be fixed, and not just simply acknowledging them"* (H4). Financial motivations also muddy these decisions, with the healthcare industry often focused on reducing costs [64]; as one participant put it, *"I get worried about what'll happen when insurance companies think there's cost savings to using these tools...they can cut corners, have more profit...given the incentives in ... the healthcare system"* (W1).

Recommended actions for practitioners:

- Clarify your research and/or impact goals before collecting data.
- Consider the current allocation of resources and where help is most needed when choosing research questions. While additional data collection can confirm well-known problems, it can also draw resources away from other important problems.
- Think through possible narratives involved in public portrayals of your work, and invest time in framing your NLP and especially LLM outputs for better public understanding of risks (e.g., hallucinations) involved in these technologies.
- Know that your measurements and predictions could contain information that could be used to incriminate the birthing person in some places and situations. Similarly, your data and measurement methods could be used for unintended purposes, like personalized advertising and the setting of insurance rates.

Learn from maternal health traditions and communities. Thousands of years of cultural traditions, healthcare practices, and tool development related to maternal health already exist [35], and more recently, grassroots communities on the internet have formed to help birthing people prepare for and work through their challenges and experiences [2, 3]. *"People have the mindset that this needs to actually replace things or that this needs to be the way in the future of addressing maternal health...diminishing all the work that's been done in the past and the positives of having people that truly understand this work."* (W2) Learning from traditions and communities can help avoid repeating mistakes made by others. *"It [(AI)] should follow the same guidelines that medical professionals do: 'Do no harm.' "* (H6)

Recommended actions for practitioners:

- Combining NLP expertise with an interdisciplinary team can help you avoid reinventing the wheel.
- Consider how NLP tools can support and learn from storytelling, which is an important language-based way for communities and generations to collectively teach and learn about maternal health, as well as a way for birthing people to process their experiences [18, 43].
- Consider how to sample training datasets without dismissing the care-seeker's perspective; weigh the risks of possible misinformation against the value of personal experiences within your specific healthcare setting.

4.2 Theme 2: Measurements

Optimize for outcomes that support the whole person. *"It's not just about the outcome, right? It's about the whole experience"* (W3). Most NLP research is centered on structured tasks with defined and accessible outcomes, e.g., predicting postpartum depression. Such tasks are important, but so are other outcomes, like achieving a positive birthing experience [41] or breastfeeding without stress [86]. As one participant says, *"it would be great if AI could tell everything from conception to birth"* (B10). Taking a more expansive view of the tasks and optimizations that could benefit maternal health stakeholders can open the door to other opportunities. By "expansive view," we do not mean that a single model must be capable of all tasks (as this may in fact be a weakness of current LLMs whose purported generalizability can lead to a lack of clear evaluation [70]). Rather, we invite NLP researchers and practitioners to be open-minded about tasks they can address with their NLP work. *"A dream AI tool should have the ability to address and develop solutions for mothers who are struggling to cope with the demands of pregnancy and overall maternal health"* (H4).

Recommended actions for practitioners:

- Curate and create text datasets with novel annotations, valuing this data work [73].
- Use social media and birthing person interviews to get a holistic overview of what outcomes birthing people care about [2].
- Run long-term user studies to support a broader range of outcomes of interest.
- Consider including subjective and/or qualitative measurements.

Protect all groups of birthing people. The healthcare system treats people differently on the basis of race, class, gender, etc. [38, 84, 103]. NLP models can mirror biases found in training datasets [16, 87], and even worse, they can over-represent and exacerbate the impacts of those biases [37, 104]. *"Everything that has the potential to benefit also has the potential to hurt"* (W4). When evaluating NLP systems, one should consider both disparities in impact (health outcomes for different groups) and disparities in treatment (whether people with the same complaint receive different treatment). Tools should *"allow providers to improve their own cultural competency or race-consciousness"* (W5), and empower *"those lacking social and financial support,"* such as young birthing people (H14).

Recommended actions for practitioners:

- Recruit birthing people from diverse groups to participate and/or give feedback on your NLP study/tool.
- Consider the role of personalization in your NLP/LLM-based system, considering both disparate impacts and disparate treatments. Sometimes people with the same complaint are not given the same treatment (e.g., pain management disparities [74]). But in other cases, applying the same treatment, without regard for individual circumstances, is also an inequity.
- When building LLM-based tools that communicate with birthing people (e.g., chatbots), evaluate across a diversity of vocabularies, accents, and languages [56].
- Beware counterfactuals as a methodological bandaid [42]. Measuring and addressing disparities are important, but simply switching tokens indicating race or gender (e.g., replacing all mentions of *she* with *he*) ignores society's multi-dimensional impacts on people living with those identities, which can lead to correlated differences in text datasets.

Hold onto the human: empathy, emotion, relationships, complexity. Emotion, empathy, and storytelling are important components of healthcare [21], and NLP tools should account for such human elements, remembering that each birth is unique and each birthing person has a unique set of circumstances, experiences, and preferences. *"Human connection is important and should be emphasized so that better quality service can be provided"* (W6); *"your own judgment and/or human compassion components, wisdom, experience [are a] part of the care"* (W7). Working with language data, as opposed to structured datasets, opens doorways to subjective tasks such as measuring empowerment [62], agency [75], and sentiment [1]. *"Birth is so complicated though that I don't see women ever replacing medical care with AI advice"* (B11).

Recommended actions for practitioners:

- Look for outliers and value what they can teach you, rather than removing them.
- A one-size-fits-all approach is probably not appropriate; prioritize personalization and/or allow the tool to be tailored by the individual birthing person or healthcare worker.
- Subjective language tasks like measuring framing [17] can illuminate aspects of healthcare experiences overlooked by non-text-based methods.
- Include qualitative methods in your study design, complementing NLP models with interviews, grounded theory [23], and methods from the social sciences.

4.3 Theme 3: Values

Include the voices of those seeking care. The voices of the most vulnerable stakeholder group, birthing people, should be included in the design and development of NLP tools. Center the principle of "nothing about us without us," most recently popularized by disability activists [19]. While it may not always be possible to include birthing people on a research or production team, it is important to integrate their perspective throughout design and deployment processes through surveys and user studies. As one participant put it, *"there needs to be community input, there needs to be representation in creating these tools"* (W8). At the same time, practitioners should be careful to engage fully with these voices, rather than using them as an ethical veneer on otherwise exploitative tools [83].

Recommended actions for practitioners:

- Where possible, include birthing people and healthcare professionals in your research team.
- Include surveys and user studies throughout your research and design process.
- Incorporate literature written by birthing people, and learn from related work about the birthing experience and birthing people's perspectives and needs.
- Be gender inclusive: do not automatically predict the gender of study subjects [44, 47], and remember to include the concerns of trans birthing people in your study design. Use gender inclusive language and follow the HCI Guidelines for Gender Equity and Inclusivity [77].

Always center the agency and autonomy of the birthing person. Maternal healthcare has an unfortunate history of abuse and disregard for the birthing person's agency (see §1.3). NLP practitioners, even when studying topics like misinformation, should not work from a perspective of skepticism about the birthing person's ability to make decisions for themselves. How NLP tools are being used to make decisions in maternal healthcare should be disclosed to the

birthing person, such that *"the person using it knows what kind of information or advice it can and cannot give"* (B2). Correspondingly, some tools should not be built [9], such as anything *"infringing on the rights of the mom or baby"* (B3).

Recommended actions for practitioners:

- Explore the construction of NLP tools that increase the agency available to the birthing person rather than making decisions for them. For example, NLP tools might be used to provide birthing people with additional decision points, resources, explanations, descriptions of what to expect, or assistance for communicating with healthcare providers.
- If NLP tools are used to make or assist in clinical decisions, this should be disclosed to birthing people, and where possible, providers should obtain direct consent from birthing people.

Respect and support your data sources. Generations of birthing people have passed down oral stories, written books, and created online content about pregnancy, labor, and the postpartum period. NLP studies and tools benefit from this knowledge and data and should give credit and be designed to avoid supplanting systems of support that are already thriving. As several workshop participants observed, ChatGPT and other similar chatbots *"don't really note their references of where they're getting the information from"* (W9). *"The principle that should guide these tools is to... have transparency for its sourcing of data"* (H2).

Recommended actions for practitioners:

- Give credit to the data sources. Use proper attribution, as this not only respects people's work but also builds trust from your users and supports auditing of your system.
- Maintain privacy. Collect only necessary data and store the data securely. Avoid perpetuating the over-surveillance of birthing people.
- Beware the "paradox of reuse" [55] in which the creation of an automated tool removes incentives for people to continue creating the training texts that power that LLMs. Encourage users to re-engage with the data source (e.g., by posting their own experiences to an online community).

5 DISCUSSION

5.1 Perceptions of risks and benefits of LLMs

Participants reported largely positive attitudes towards NLP tools with important caveats. For example, workshop participants expressed high hopes about the development and adoption of NLP tools in healthcare but unanimously agreed that NLP tools for maternal health should always be designed to be assistive rather than autonomous (*holding on to the human*). Settings in which NLP tools were viewed to be particularly useful to healthcare providers included: (i) reducing administrative burden, including better/faster communication and coordination of care across medical departments, and (ii) improving medical education and training, including assisting staff in developing better cultural sensitivity. On the other hand, tools for decision-making settings were viewed negatively. These results contrast with current trends in NLP research (§1), which tend towards surveillance of birthing people and tasks like risk prediction.

Birthing people viewed NLP tools very positively, especially for providing information and recommendations (*centering their agency and autonomy*). For example, participants wrote that they wished they had the chatbot during

their pregnancy: *"I wish it [(AI)] had been around when my son was a newborn so I could interact with it during late night feedings. One, to give me something to do, and two, to make me feel like I wasn't alone"* (B12). In particular, birthing people emphasized how comforting it would be to have a fast and convenient resource to assuage fears about whether what they were experiencing was normal: *"Just to have something there to ask questions to when I am not sure as to what is happening or when I need a quick answer"* (B9). Some participants cautioned that such tools should provide additional context to help patients understand whether the suggestions or advice are applicable to their specific situation due to differing levels of medical understanding.

These positive impressions need to be weighed carefully against research demonstrating potential harms. On one hand, both birthing people and clinicians are desperately in need of support, as e.g., postpartum depression rates [88] and clinician burnout rates [30] demonstrate. Challenges around healthcare costs, inaccessibility, and community distrust of the medical establishment [26] support the positive views of our participants towards NLP and LLM-based tools. But on the other hand, the current capabilities of these tools are limited and their vulnerabilities well-documented [98], and for some challenges, non-technical solutions already exist whose implementation might be further delayed by allocating resources to NLP technologies. Researchers and practitioners should be ready to abandon NLP technologies for maternal health if there are not clear benefits.

5.2 LLMs as part of a larger ecosystem to meet information needs

In their free-text responses, about half of the birthing people and two healthcare workers compared their experience with the chatbot to their past experiences with internet search engines. Most framed the comparison favorably, writing *"Often times people will google questions and try to sift through all the search results to find the applicable information. AI could make that a much more efficient process"* (H9) and *"It would be nice to be able to type in worries and fears to an AI bot and get accurate answers instead of going down rabbit trails on search engines that leave you more concerned"* (B11). On the other hand, one participant noted dangers to LLMs similar to known dangers of popular healthcare websites: *"People already diagnose themselves on WebMD. Providing more tools can be dangerous"* (B4). These participants viewed LLMs as one more tool in an existing information ecosystem, exacerbating or ameliorating many of the same risks. This view is complicated by recent work by Shah and Bender [79] that emphasizes the risks of using LLMs as replacements for traditional search interfaces, both to their users (via hallucinations) and to entire information ecosystems that can be polluted by incorrect and ungrounded outputs of LLMs; this work calls for slower research that prioritizes answering fundamental questions about evaluation, costs, etc. before deploying LLMs as information retrieval systems.

5.3 Comparison to prior ethics guidelines

Multiple guidelines for machine learning applied to healthcare applications already exist [22, 54, 68, 78, 100]; what does our study add to these? (i) Our study is the first to focus on maternal health; this is a critical area with infamous historical abuses that warrants its own ethical studies. Our work is also novel in (ii) our focus on new NLP technologies like LLMs and (iii) grounding our guiding principles in perceptions from multiple stakeholder groups. Finally, (iv) the principles we uncover are distinct from those in prior work. While some of the themes overlap with prior work (e.g., problem formulation [78]), many of our themes (e.g., power dynamics in the care team, political implications of measurements) have not been raised in prior guidelines, which have often been focused on machine learning vulnerabilities rather than the social systems surrounding these technologies. Compared to the two studies closest to ours by Sendak et al. [78] and Petti et al. [68], which focus on ethical guidelines for sepsis and Alzheimer's disease, respectively, our focus on

maternal health uncovers principles specific to NLP and LLMs and emphasizes communities of birthing people and other care-seekers: their knowledge and tool development, their autonomy, and their historical and ongoing vulnerabilities.

5.4 Generalization to other healthcare settings

Maternal health served as a representative health case study through which we could apply a focused lens to the use of NLP and LLMs for healthcare. Maternal health is a pressure cooker of modern biases, historical injustices, misinformation, care team imbalances, and other challenges that appear in many other healthcare contexts. Related healthcare settings include resource shortages in primary care [14], individuals seeking health information to diagnose symptoms [4], and racial considerations in disease severity models [65]. In each case, our guiding principles can help practitioners think through whether and how to incorporate NLP tools into new healthcare settings.

5.5 Limitations

Our survey respondents are not representative of all birthing people worldwide. The birthing people in our study were from the U.S., spoke English, and lack representation among Asian and Pacific Islander groups. We used ChatGPT-3.5 for the demonstration, since its rate limits fit the constraints of the live workshop, though the model is regularly updated and results may not be replicable. We highlight the need for longitudinal studies of LLM usage in the healthcare space, to illuminate when to use LLMs in the clinical tool development process and also the compounding effects of AI and LLMs on birthing people over time.

Each of our participant cohorts had a unique selection of values, forming individual priority "signatures" (Figure 2). Differences in results between our participant cohorts are likely driven by at least three factors, though we cannot make any conclusive causal claims: (i) the groups represent different stakeholders; (ii) the workshop participants were primed by their multi-hour exposure to lectures about NLP biases; and (iii) the workshop participants opted into the workshop, which was themed around "maternal health equity," while the Prolific participants were solicited online and paid for their work. We also observed a notable difference in prior AI usage and general perceptions of AI between the workshop participants and the two groups of healthcare workers and birthing people recruited from Prolific. Workshop participants on average were much less familiar with AI and held less positive perceptions of AI (Figures 3 and 4 in the Appendix).

It is important to engage with multiple groups in different settings and integrating their viewpoints, not by enforcing agreement but by surfacing the inherent tensions and different viewpoints that exist in the perceptions of these technologies.

6 CONCLUSION

In consultation with healthcare professionals, birthing people, and other stakeholders, we developed a set of guiding principles for the use of NLP and LLMs in maternal healthcare. This work serves as a guide to researchers and practitioners broadly on how to engage with affected peoples in healthcare contexts. We hope that researchers and clinicians will work with the people most affected—the people seeking care—and will read and follow the principles derived by our methods. Machine learning researchers should also explore how their tools can actively broadcast these principles or how model development pipelines can be adapted to better meet the needs of all stakeholders.

REFERENCES

[1] Tanveer Ali, David Schramm, Marina Sokolova, and Diana Inkpen. 2013. Can I Hear You? Sentiment Analysis on Medical Forums. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, Ruslan Mitkov and Jong C. Park (Eds.). Asian Federation of Natural Language Processing, Nagoya, Japan, 667–673. https://aclanthology.org/I13-1077

[2] Nazanin Andalibi and Patricia Garcia. 2021. Sensemaking and Coping After Pregnancy Loss: The Seeking and Disruption of Emotional Validation Online. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 127 (apr 2021), 32 pages. https://doi.org/10.1145/3449201

[3] Maria Antoniak, David Mimno, and Karen Levy. 2019. Narrative Paths and Negotiation of Power in Birth Stories. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 88 (nov 2019), 27 pages. https://doi.org/10.1145/3359190

[4] John W. Ayers, Adam Poliak, Mark Dredze, Eric C. Leas, Zechariah Zhu, Jessica B. Kelley, Dennis J. Faix, Aaron M. Goodman, Christopher A. Longhurst, Michael Hogarth, and Davey M. Smith. 2023. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Internal Medicine* 183, 6 (06 2023), 589–596. https://doi.org/10.1001/jamainternmed.2023.1838 arXiv:https://jamanetwork.com/journals/jamainternalmedicine/articlepdf/2804309/jamainternal_ayers_2023_oi_230030_1685974538.66672.pdf

[5] 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 (06 2023), e2317517–e2317517. https://doi.org/10.1001/jamanetworkopen.2023.17517 arXiv:https://jamanetwork.com/journals/jamanetworkopen/articlepdf/2805756/ayers_2023_ld_230091_1685466899.60716.pdf

[6] Oliver Baclic, Matthew C. Tunis, Kelsey Young, Coraline Doan, Howard Swerdfeger, and Justin Schonfeld. 2020. Challenges and opportunities for public health made possible by advances in natural language processing. *Canada communicable disease report = Releve des maladies transmissibles au Canada* 46 6 (2020), 161–168. https://api.semanticscholar.org/CorpusID:219904380

[7] Aparna Balagopalan, Haoran Zhang, Kimia Hamidieh, Thomas Hartvigsen, Frank Rudzicz, and Marzyeh Ghassemi. 2022. The Road to Explainability is Paved with Bias: Measuring the Fairness of Explanations. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1194–1206. https://doi.org/10.1145/3531146.3533179

[8] Amanda Baughan, Xuezhi Wang, Ariel Liu, Allison Mercurio, Jilin Chen, and Xiao Ma. 2023. A Mixed-Methods Approach to Understanding User Trust after Voice Assistant Failures. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 7, 16 pages. https://doi.org/10.1145/3544548.3581152

[9] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922

[10] Ruha Benjamin. 2022. A Better Birth Is Possible. [Online; posted 22-September-2022].

[11] CJ Berg, WM Callaghan, C Syverson, and Z Henderson. 2010. Pregnancy-related mortality in the United States, 1998 to 2005. *Obstetrics & Gynecology* 116, 6 (2010), 1302.

[12] Cynthia J Berg, Margaret A Harper, Samuel M Atkinson, Elizabeth A Bell, Haywood L Brown, Marvin L Hage, Avick G Mitra, Kenneth J Moise, and William M Callaghan. 2005. Preventability of pregnancy-related deaths: results of a state-wide review. *Obstetrics & Gynecology* 106, 6 (2005), 1228–1234.

[13] William Boag, Hassan Kané, Saumya Rawat, Jesse Wei, and Alexander Goehler. 2021. A Pilot Study in Surveying Clinical Judgments to Evaluate Radiology Report Generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 458–465. https://doi.org/10.1145/3442188.3445909

[14] Thomas S Bodenheimer and Mark D Smith. 2013. Primary care: proposed solutions to the physician shortage without training more physicians. *Health Affairs* 32, 11 (2013), 1881–1886.

[15] Phyllis L. Brodsky. 2008. Where Have All the Midwives Gone? *Journal of Perinatal Education* 17 (2008), 48 – 51. https://api.semanticscholar.org/CorpusID:39081236

[16] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2016. Semantics derived automatically from language corpora contain human-like biases. *Science* 356 (2016), 183 – 186. https://api.semanticscholar.org/CorpusID:23163324

[17] Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The Media Frames Corpus: Annotations of Frames Across Issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Chengqing Zong and Michael Strube (Eds.). Association for Computational Linguistics, Beijing, China, 438–444. https://doi.org/10.3115/v1/P15-2072

[18] Anna Carson, Cathy Chabot, Devon Greyson, Kate Shannon, Putu Duff, and Jean Shoveller. 2017. A narrative analysis of the birth stories of early-age mothers. *Sociology of Health & Illness* 39, 6 (2017), 816–831.

[19] James I. Charlton. 1998. Nothing About Us Without Us: Disability Oppression and Empowerment. https://api.semanticscholar.org/CorpusID:153691995

[20] Kathy Charmaz. 2014. *Constructing Grounded Theory*. Sage.

[21] Rita Charon. 2001. Narrative Medicine: A Model for Empathy, Reflection, Profession, and Trust. *JAMA* 286, 15 (10 2001), 1897–1902. https://doi.org/10.1001/jama.286.15.1897 arXiv:https://jamanetwork.com/journals/jama/articlepdf/194300/jrp10002.pdf

-

[22] Irene Y. Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. 2021. Ethical Machine Learning in Healthcare. *Annual Review of Biomedical Data Science* 4, 1 (2021), 123–144. https://doi.org/10.1146/annurev-biodatasci-092820-114757 arXiv:https://doi.org/10.1146/annurev-biodatasci-092820-114757

[23] Juliet M. Corbin and Anselm Strauss. 1990. Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology* 13 (1990), 3–21. https://api.semanticscholar.org/CorpusID:12711240

[24] Andreea A Creanga, Cynthia J. Berg, Jean Y. Ko, Sherry L Farr, Van T Tong, Francesca Bruce, and William M. Callaghan. 2014. Maternal mortality and morbidity in the United States: where are we now? *Journal of Women's Health* 23 1 (2014), 3–9.

[25] Andreea A Creanga, Cynthia J Berg, Carla Syverson, Kristi Seed, F Carol Bruce, and William M Callaghan. 2015. Pregnancy-related mortality in the United States, 2006-2010. *Obstet. Gynecol.* 125, 1 (Jan. 2015), 5–12.

[26] Adolfo G Cuevas, Kerth O'Brien, and Somnath Saha. 2016. African American experiences in healthcare:"I always feel like I'm getting skipped over". *Health Psychology* 35, 9 (2016), 987.

[27] Eugene R Declercq, Howard J Cabral, Chia-Ling Liu, Ndidiamaka Amutah-Onukagha, Audra Meadows, Xiaohui Cui, and Hafsatou Diop. 2023. Prior Hospitalization, Severe Maternal Morbidity, and Pregnancy-Associated Deaths in Massachusetts From 2002 to 2019. *Obstetrics & Gynecology* 142, 6 (2023), 1423–1430.

[28] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (Boston, MA, USA) *(EAAMO '23)*. Association for Computing Machinery, New York, NY, USA, Article 37, 23 pages. https://doi.org/10.1145/3617694.3623261

[29] Catherine D'Ignazio and Lauren F. Klein. 2023. *Data Feminism.* MIT Press.

[30] Victor J Dzau, Darrell G Kirch, Thomas J Nasca, et al. 2018. To care is human—collectively confronting the clinician-burnout crisis. *N Engl J Med* 378, 4 (2018), 312–314.

[31] Scott L Fleming, Alejandro Lozano, William J Haberkorn, Jenelle A Jindal, Eduardo P Reis, Rahul Thapa, Louis Blankemeier, Julian Z Genkins, Ethan Steinberg, Ashwin Nayak, et al. 2023. MedAlign: A Clinician-Generated Dataset for Instruction Following with Electronic Medical Records. *arXiv preprint arXiv:2308.14089* (2023).

[32] Riccardo Fogliato, Shreya Chappidi, Matthew Lungren, Paul Fisher, Diane Wilson, Michael Fitzke, Mark Parkinson, Eric Horvitz, Kori Inkpen, and Besmira Nushi. 2022. Who Goes First? Influences of Human-AI Workflow on Decision Making in Clinical Imaging. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1362–1374. https://doi.org/10.1145/3531146.3533193

[33] Marzyeh Ghassemi and Shakir Mohamed. 2022. Machine learning and health need better values. *NPJ Digital Medicine* 5 (2022).

[34] Aziliz Le Glaz, Yannis Haralambous, Deok-Hee Kim-Dufor, Philippe Lenca, Romain Billot, Taylor C. Ryan, Jonathan J. Marsh, Jordan Devylder, Michel Walter, Sofian Berrouiguet, and Christophe Lemey. 2019. Machine Learning and Natural Language Processing in Mental Health: Systematic Review. *Journal of Medical Internet Research* 23 (2019). https://api.semanticscholar.org/CorpusID:233719081

[35] Keisha Goode and Barbara Katz Rothman. 2017. African-American midwifery, a history and a lament. *American Journal of Economics and Sociology* 76, 1 (2017), 65–94.

[36] Heath Goodrum, Meghana Gudala, Ankita Misra, and Kirk Roberts. 2019. Extraction of Lactation Frames from Drug Labels and LactMed. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, Dina Demner-Fushman, Kevin Bretonnel Cohen, Sophia Ananiadou, and Junichi Tsujii (Eds.). Association for Computational Linguistics, Florence, Italy, 191–200. https://doi.org/10.18653/v1/W19-5020

[37] Melissa Hall, Laurens van der Maaten, Laura Gustafson, and Aaron B. Adcock. 2022. A Systematic Study of Bias Amplification. *ArXiv* abs/2201.11706 (2022). https://api.semanticscholar.org/CorpusID:246294816

[38] Monica Webb Hooper, Anna María Nápoles, and Eliseo J Perez-stable. 2020. COVID-19 and Racial/Ethnic Disparities. *JAMA* (2020). https://api.semanticscholar.org/CorpusID:218584885

[39] Nanna Inie, Jeanette Falk, and Steve Tanimoto. 2023. Designing Participatory AI: Creative Professionals' Worries and Expectations about Generative AI. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 82, 8 pages. https://doi.org/10.1145/3544549.3585657

[40] Maurice Jakesch, Zana Buçinca, Saleema Amershi, and Alexandra Olteanu. 2022. How Different Groups Prioritize Ethical Values for Responsible AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 310–323. https://doi.org/10.1145/3531146.3533097

[41] Annika Karlström, Astrid Nystedt, and Ingegerd Hildingsson. 2015. The meaning of a very positive birth experience: focus groups discussions with women. *BMC Pregnancy and Childbirth* 15 (2015). https://api.semanticscholar.org/CorpusID:18330640

[42] Atoosa Kasirzadeh and Andrew Smart. 2021. The Use and Misuse of Counterfactuals in Ethical Machine Learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 228–236. https://doi.org/10.1145/3442188.3445886

[43] Lesley Kay, Soo Downe, Gill Thomson, and Kenny Finlayson. 2017. Engaging with birth stories in pregnancy: a hermeneutic phenomenological study of women's experiences across two generations. *BMC Pregnancy and Childbirth* 17, 1 (2017), 283.

[44] Os Keyes. 2017. Stop Mapping Names to Gender. https://ironholds.org/names-gender/. Accessed: 2021-05-26.

[45] Jee Young Kim, William Boag, Freya Gulamali, Alifia Hasan, Henry David Jeffry Hogg, Mark Lifson, Deirdre Mulligan, Manesh Patel, Inioluwa Deborah Raji, Ajai Sehgal, et al. 2023. Organizational Governance of Emerging Technologies: AI Adoption in Healthcare. In *Proceedings of the 2023*

ACM Conference on Fairness, Accountability, and Transparency. 1396–1417.

[46] Rebecca Knowles, Mark Dredze, Kathleen Evans, Elyse Lasser, Tom Richards, Jonathan Weiner, and Hadi Kharrazi. 2014. High risk pregnancy prediction from clinical text. In *Proceeding of the NIPS Workshop on Machine Learning for Clinical Data Analysis*. 1–4.

[47] Brian Larson. 2017. Gender as a Variable in Natural-Language Processing: Ethical Considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Association for Computational Linguistics, Valencia, Spain, 1–11. https://doi.org/10.18653/v1/W17-1601

[48] María Laura Ramírez Galleguillos and Aykut Coşkun. 2020. How Do I Matter? A Review of the Participatory Design Practice with Less Privileged Participants. In *Proceedings of the 16th Participatory Design Conference 2020 - Participation(s) Otherwise - Volume 1* (Manizales, Colombia) *(PDC '20)*. Association for Computing Machinery, New York, NY, USA, 137–147. https://doi.org/10.1145/3385010.3385018

[49] Lung-Hao Lee, Man-Chen Hung, Chien-Huan Lu, Chang-Hao Chen, Po-Lei Lee, and Kuo-Kai Shyu. 2021. Classification of Tweets Self-reporting Adverse Pregnancy Outcomes and Potential COVID-19 Cases Using RoBERTa Transformers. In *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, Arjun Magge, Ari Klein, Antonio Miranda-Escalada, Mohammed Ali Al-garadi, Ilseyar Alimova, Zulfat Miftahutdinov, Eulalia Farre-Maduell, Salvador Lima Lopez, Ivan Flores, Karen O'Connor, Davy Weissenbacher, Elena Tutubalina, Abeed Sarker, Juan M Banda, Martin Krallinger, and Graciela Gonzalez-Hernandez (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 98–101. https://doi.org/10.18653/v1/2021.smm4h-1.18

[50] Peter Lee, Carey Goldberg, and Isaac Kohane. 2023. *The AI revolution in medicine: GPT-4 and beyond*. Pearson.

[51] I. S. L. Loudon. 2008. General practitioners and obstetrics: a brief history. *Journal of the Royal Society of Medicine* 101 (2008), 531 – 535. https://api.semanticscholar.org/CorpusID:32952274

[52] Marian F MacDorman, Eugene Declercq, Howard Cabral, and Christine Morton. 2016. Recent Increases in the U.S. Maternal Mortality Rate: Disentangling Trends From Measurement Issues. *Obstet. Gynecol.* 128, 3 (Sept. 2016), 447–455.

[53] Nina Martin and Renee Montagne. 2017. Nothing Protects Black Women From Dying in Pregnancy and Childbirth. *ProPublica* (2017).

[54] Melissa Mccradden, Oluwadara Odusi, Shalmali Joshi, Ismail Akrout, Kagiso Ndlovu, Ben Glocker, Gabriel Maicas, Xiaoxuan Liu, Mjaye Mazwi, Tee Garnett, Lauren Oakden-Rayner, Myrtede Alfred, Irvine Sihlahla, Oswa Shafei, and Anna Goldenberg. 2023. What's fair is… fair? Presenting JustEFAB, an ethical framework for operationalizing medical ethics and social justice in the integration of clinical machine learning: JustEFAB. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1505–1519. https://doi.org/10.1145/3593013.3594096

[55] Connor McMahon, Isaac L. Johnson, and Brent J. Hecht. 2017. The Substantial Interdependence of Wikipedia and Google: A Case Study on the Relationship Between Peer Production Communities and Information Technologies. In *International Conference on Web and Social Media*. https://api.semanticscholar.org/CorpusID:810239

[56] Nikita Mehandru, Samantha Robertson, and Niloufar Salehi. 2022. Reliable and Safe Use of Machine Translation in Medical Settings. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 2016–2025. https://doi.org/10.1145/3531146.3533244

[57] Hossein Mohammadhassanzadeh, Ingrid Sketris, Robyn Traynor, Susan Alexander, Brandace Winquist, Samuel Alan Stewart, et al. 2020. Using natural language processing to examine the uptake, content, and readability of media coverage of a pan-canadian drug safety research project: Cross-sectional observational study. *JMIR Formative Research* 4, 1 (2020), e13296.

[58] Michael J Muller and Sarah Kuhn. 1993. Participatory design. *Commun. ACM* 36, 6 (1993), 24–28.

[59] Aakanksha Naik, Sravanthi Parasa, Sergey Feldman, Lucy Lu Wang, and Tom Hope. 2022. Literature-Augmented Clinical Outcome Prediction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 438–453. https://doi.org/10.18653/v1/2022.findings-naacl.33

[60] Gary S Nickell and John N Pinto. 1986. The computer attitude scale. *Computers in Human Behavior* 2, 4 (1986), 301–306.

[61] Azadeh Nikfarjam, A. Sarker, Karen O'Connor, Rachel E. Ginn, and Graciela Gonzalez-Hernandez. 2015. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association : JAMIA* 22 (2015), 671 – 681. https://api.semanticscholar.org/CorpusID:8319232

[62] Lucille Njoo, Chan Park, Octavia Stappart, Marvin Thielk, Yi Chu, and Yulia Tsvetkov. 2023. TalkUp: Paving the Way for Understanding Empowering Language. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 9334–9354. https://doi.org/10.18653/v1/2023.findings-emnlp.625

[63] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of GPT-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375* (2023).

[64] Ziad Obermeyer and Sendhil Mullainathan. 2019. Dissecting Racial Bias in an Algorithm that Guides Health Decisions for 70 Million People. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 89. https://doi.org/10.1145/3287560.3287593

[65] Jesutofunmi A Omiye, Jenna C Lester, Simon Spichak, Veronica Rotemberg, and Roxana Daneshjou. 2023. Large language models propagate race-based medicine. *NPJ Digital Medicine* 6, 1 (2023), 195.

[66] Siru Ouyang, Shuohang Wang, Yang Liu, Ming Zhong, Yizhu Jiao, Dan Iter, Reid Pryzant, Chenguang Zhu, Heng Ji, and Jiawei Han. 2023. The Shifted and The Overlooked: A Task-oriented Investigation of User-GPT Interactions. arXiv:2310.12418 [cs.CL]

[67] Cecilia Panigutti, Alan Perotti, and Dino Pedreschi. 2020. Doctor XAI: an ontology-based approach to black-box sequential data classification explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for

Computing Machinery, New York, NY, USA, 629–639. https://doi.org/10.1145/3351095.3372855

[68] Ulla Petti, Rune Nyrup, Jeffrey M. Skopek, and Anna Korhonen. 2023. Ethical considerations in the early detection of Alzheimer's disease using speech and AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1062–1075. https://doi.org/10.1145/3593013.3594063

[69] Raphael Poulain, Mirza Farhan Bin Tarek, and Rahmatollah Beheshti. 2023. Improving Fairness in AI Models on Electronic Health Records: The Case for Federated Learning Methods. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1599–1608. https://doi.org/10.1145/3593013.3594102

[70] Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. 2021. AI and the Everything in the Whole Wide World Benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, J. Vanschoren and S. Yeung (Eds.), Vol. 1. Curran. https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/084b6fbb10729ed4da8c3d3f5a3ae7c9-Paper-round2.pdf

[71] Alvin Rajkomar, Michaela Hardt, Michael D. Howell, Greg Corrado, and Marshall H. Chin. 2018. Ensuring Fairness in Machine Learning to Advance Health Equity. *Annals of Internal Medicine* 169, 12 (2018), 866–872.

[72] Negar Rostamzadeh, Diana Mincu, Subhrajit Roy, Andrew Smart, Lauren Wilcox, Mahima Pushkarna, Jessica Schrouff, Razvan Amironesei, Nyalleng Moorosi, and Katherine Heller. 2022. Healthsheet: Development of a Transparency Artifact for Health Datasets. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1943–1961. https://doi.org/10.1145/3531146.3533239

[73] 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* (<conf-loc>, <city>Yokohama</city>, <country>Japan</country>, </conf-loc>) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 39, 15 pages. https://doi.org/10.1145/3411764.3445518

[74] Anke Samulowitz, Ida Gremyr, Erik Masao Eriksson, and Gunnel Hensing. 2018. "Brave Men" and "Emotional Women": A Theory-Guided Literature Review on Gender Bias in Health Care and Gendered Norms towards Patients with Chronic Pain. *Pain Research & Management* 2018 (2018). https://api.semanticscholar.org/CorpusID:5034523

[75] Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation Frames of Power and Agency in Modern Films. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Martha Palmer, Rebecca Hwa, and Sebastian Riedel (Eds.). Association for Computational Linguistics, Copenhagen, Denmark, 2329–2334. https://doi.org/10.18653/v1/D17-1247

[76] Guergana K. Savova, James J. Masanz, Philip V. Ogren, Jiaping Zheng, Sunghwan Sohn, Karin Kipper Schuler, and Christopher G. Chute. 2010. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association : JAMIA* 17 5 (2010), 507–13. https://api.semanticscholar.org/CorpusID:564263

[77] Morgan Klaus Scheuerman, Katta Spiel, Oliver L Haimson, Foad Hamidi, and Stacy M Branham. 2020. HCI guidelines for gender equity and inclusivity. (2020).

[78] Mark Sendak, Madeleine Clare Elish, Michael Gao, Joseph Futoma, William Ratliff, Marshall Nichols, Armando Bedoya, Suresh Balu, and Cara O'Brien. 2020. "The human body is a black box": supporting clinical decision-making with deep learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 99–109. https://doi.org/10.1145/3351095.3372827

[79] Chirag Shah and Emily M Bender. 2023. Envisioning Information Access Systems: What Makes for Good Tools and a Healthy Web? (2023).

[80] Farah E. Shamout, Tingting Zhu, and David A. Clifton. 2020. Machine Learning for Clinical Outcome Prediction. *IEEE Reviews in Biomedical Engineering* 14 (2020), 116–126. https://api.semanticscholar.org/CorpusID:221068450

[81] Yuqi Si, Jingqi Wang, Hua Xu, and Kirk Roberts. 2019. Enhancing Clinical Concept Extraction with Contextual Embedding. *Journal of the American Medical Informatics Association : JAMIA* (2019). https://api.semanticscholar.org/CorpusID:67856085

[82] Venkatesh Sivaraman, Leigh A Bukowski, Joel Levin, Jeremy M. Kahn, and Adam Perer. 2023. Ignore, Trust, or Negotiate: Understanding Clinician Acceptance of AI-Based Treatment Recommendations in Health Care. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 754, 18 pages. https://doi.org/10.1145/3544548.3581075

[83] Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. 2022. Participation Is not a Design Fix for Machine Learning. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (<conf-loc>, <city>Arlington</city>, <state>VA</state>, <country>USA</country>, </conf-loc>) *(EAAMO '22)*. Association for Computing Machinery, New York, NY, USA, Article 1, 6 pages. https://doi.org/10.1145/3551624.3555285

[84] Brian D. Smedley, Adrienne Y. Stith, and Alan R. Nelson. 2003. Unequal treatment: confronting racial and ethnic disparities in health care. *Journal of The National Medical Association* 94 (2003), 666. https://api.semanticscholar.org/CorpusID:37575839

[85] Taylor Sorensen, Liwei Jiang, Jena Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, et al. 2023. Value Kaleidoscope: Engaging AI with Pluralistic Human Values, Rights, and Duties. *arXiv preprint arXiv:2309.00779* (2023).

[86] Mariz Spannhake, Charlotte Jansen, Tatiana Görig, and Katharina Diehl. 2021. "It Is a Very Emotional Topic for Me"—Managing Breastfeeding Problems among German Mothers: A Qualitative Approach. *Healthcare* 9 (2021). https://api.semanticscholar.org/CorpusID:239472182

[87] Ryan Steed, Swetasudha Panda, Ari Kobren, and Michael Wick. 2022. Upstream Mitigation Is *Not* All You Need: Testing the Bias Transfer Hypothesis in Pre-Trained Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume*

1: Long Papers), Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 3524–3542. https://doi.org/10.18653/v1/2022.acl-long.247

[88] Donna E Stewart and Simone Vigod. 2016. Postpartum Depression. *N. Engl. J. Med.* 375, 22 (Dec. 2016), 2177–2186.

[89] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature Medicine* 29, 8 (2023), 1930–1940.

[90] Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Yong Cheng, Le Hou, Albert Webson, Kavita Kulkarni, S Sara Mahdavi, Christopher Semturs, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S Corrado, Yossi Matias, Alan Karthikesalingam, and Vivek Natarajan. 2024. Towards Conversational Diagnostic AI. arXiv:2401.05654 [cs.AI]

[91] Brock E.W. Turner. 2023. Epic, Microsoft bring GPT-4 to EHRs. https://www.modernhealthcare.com/digital-health/himss-2023-epic-microsoft-bring-openais-gpt-4-ehrs. [Accessed 05-11-2023].

[92] Daiju Ueda, Shannon Walston, Toshimasa Matsumoto, Ryo Deguchi, Hiroyuki Tatekawa, and Yukio Miki. 2023. Evaluating GPT-4-based ChatGPT's Clinical Potential on the NEJM Quiz. *medRxiv* (2023), 2023–05.

[93] Raju Vaishya, Anoop Misra, and Abhishek Vaish. 2023. ChatGPT: Is this version good for healthcare and research? *Diabetes & Metabolic Syndrome: Clinical Research & Reviews* 17, 4 (2023), 102744.

[94] Luz Itzel Valdeolivar-Hernandez, Maria Eugenia Flores Quijano, Juan Carlos Echeverria-Arjonilla, Jorge Perez-Gonzalez, and Omar Piña-Ramirez. 2023. Towards breastfeeding self-efficacy and postpartum depression estimation based on analysis of free-speech interviews through natural language processing. In *18th International Symposium on Medical Information Processing and Analysis*, Jorge Brieva, Pamela Guevara, Natasha Lepore, Marius G. Linguraru, Letícia Rittner, and Eduardo Romero Castro M.D. (Eds.), Vol. 12567. International Society for Optics and Photonics, SPIE, 125670T. https://doi.org/10.1117/12.2669883

[95] Betty van Aken, Jens-Michalis Papaioannou, Manuel Mayrdorfer, Klemens Budde, Felix Gers, and Alexander Loeser. 2021. Clinical Outcome Prediction from Admission Notes using Self-Supervised Knowledge Integration. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (Eds.). Association for Computational Linguistics, Online, 881–893. https://doi.org/10.18653/v1/2021.eacl-main.75

[96] Saraswathi Vedam, Kathrin Stoll, Tanya Khemet Taiwo, Nicholas Rubashkin, Melissa Cheyney, Nan Strauss, Monica McLemore, Micaela Cadena, Elizabeth Nethery, Eleanor Rushton, et al. 2019. The Giving Voice to Mothers study: Inequity and mistreatment during pregnancy and childbirth in the United States. *Reproductive Health* 16 (2019), 1–18.

[97] Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, and Hongfang Liu. 2018. Clinical information extraction applications: A literature review. *Journal of biomedical informatics* 77 (2018), 34–49. https://api.semanticscholar.org/CorpusID:3632923

[98] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from Language Models. arXiv:2112.04359 [cs.CL]

[99] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of Risks posed by Language Models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 214–229. https://doi.org/10.1145/3531146.3533088

[100] Jenna Wiens, Suchi Saria, Mark P. Sendak, Marzyeh Ghassemi, Vincent X. Liu, Finale Doshi-Velez, Kenneth Jung, Katherine A. Heller, David C. Kale, Mohammed Saeed, Pilar N. Ossorio, Sonoo Thadaney-Israni, and Anna Goldenberg. 2019. Do no harm: a roadmap for responsible machine learning for health care. *Nature Medicine* 25 (2019), 1337 – 1340.

[101] Michael Wornow, Yizhe Xu, Rahul Thapa, Birju Patel, Ethan Steinberg, Scott Fleming, Michael A Pfeffer, Jason Fries, and Nigam H Shah. 2023. The shaky foundations of large language models and foundation models for electronic health records. *npj Digital Medicine* 6, 1 (2023), 135.

[102] Yuxin Xiao, Shulammite Lim, Tom Joseph Pollard, and Marzyeh Ghassemi. 2023. In the Name of Fairness: Assessing the Bias in Clinical Record De-identification. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 123–137. https://doi.org/10.1145/3593013.3593982

[103] Valentina A. Zavala, Paige M Bracci, John M. Carethers, Luis G. Carvajal-Carmona, Nicole B. Coggins, Marcia Cruz-Correa, Melissa B. Davis, Adam J. de Smith, Julie Dutil, Jane C. Figueiredo, Rena K. Fox, Kristi D. Graves, Scarlett Lin Gomez, Andrea Sabina Llera, Susan L. Neuhausen, Lisa A Newman, Tung T. Nguyen, Julie R. Palmer, Nynikka R. Palmer, Eliseo J Perez-stable, Sorbarikor Piawah, Erik J. Rodriquez, María Carolina Sanabria-Salas, Stephanie L. Schmit, Silvia J Serrano-Gómez, Mariana C. Stern, Jeffrey N. Weitzel, Jun J. Yang, Jovanny Zabaleta, Elad Ziv, and Laura Fejerman. 2020. Cancer health disparities in racial/ethnic minorities in the United States. *British Journal of Cancer* 124 (2020), 315 – 332. https://api.semanticscholar.org/CorpusID:221570622

[104] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In *Conference on Empirical Methods in Natural Language Processing*. https://api.semanticscholar.org/CorpusID:1389483

-

A EXTENDED SURVEY RESULTS

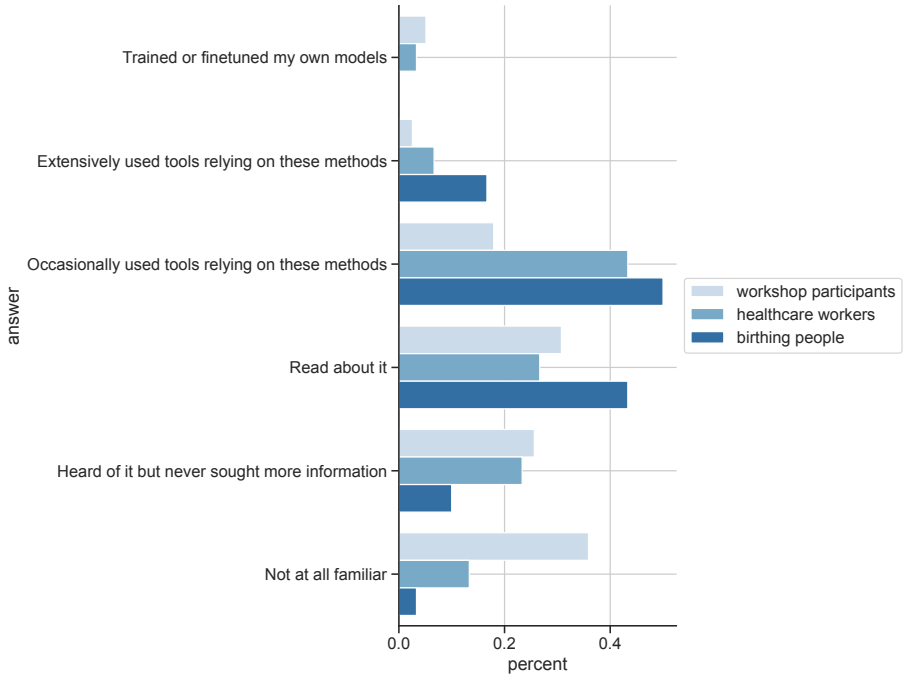

Fig. 3. Answers to the survey question *[Before this workshop,] How familiar were/are you with NLP, machine learning, and/or AI?* Overall, the workshop participants were less familiar with these topics than the healthcare workers and birthing people who were recruited via the Prolific platform.

Alt Text: A bar plot showing the answers to the survey question about familiarity with NLP/ML/AI, with responses broken apart by participant cohort.

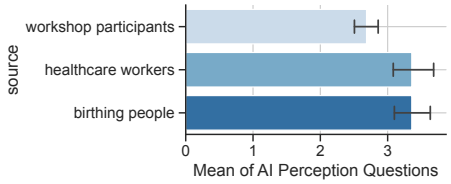

Fig. 4. Answers to the four survey questions about general perceptions of AI. Overall, the workshop participants less frequently reported positive perceptions than the healthcare workers and birthing people who were recruited via the Prolific platform.

Alt Text: A bar plot showing the averaged answers to the survey questions about general perceptions of AI, with responses broken apart by participant cohort.

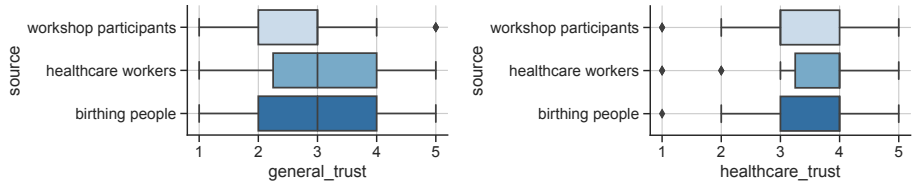

Fig. 5. Answers to the two survey questions about trust. Overall, the healthcare workers were more trusting of healthcare providers.

Alt Text: Two box plots showing the answers to the survey questions about trust, with responses broken apart by participant cohort.

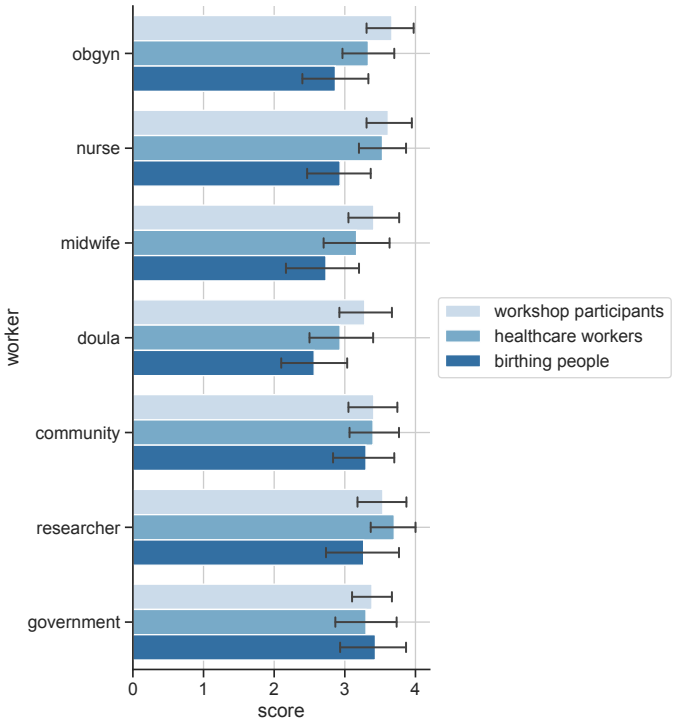

Fig. 6. Answers to the survey questions about who would benefit from an AI/NLP chatbot in their work, with the 5-point Likert scale score shown on the x-axis.

Alt Text: A bar plot showing the answers to the survey question about who would benefit from an AI/NLP chatbot, with responses broken apart by participant cohort.

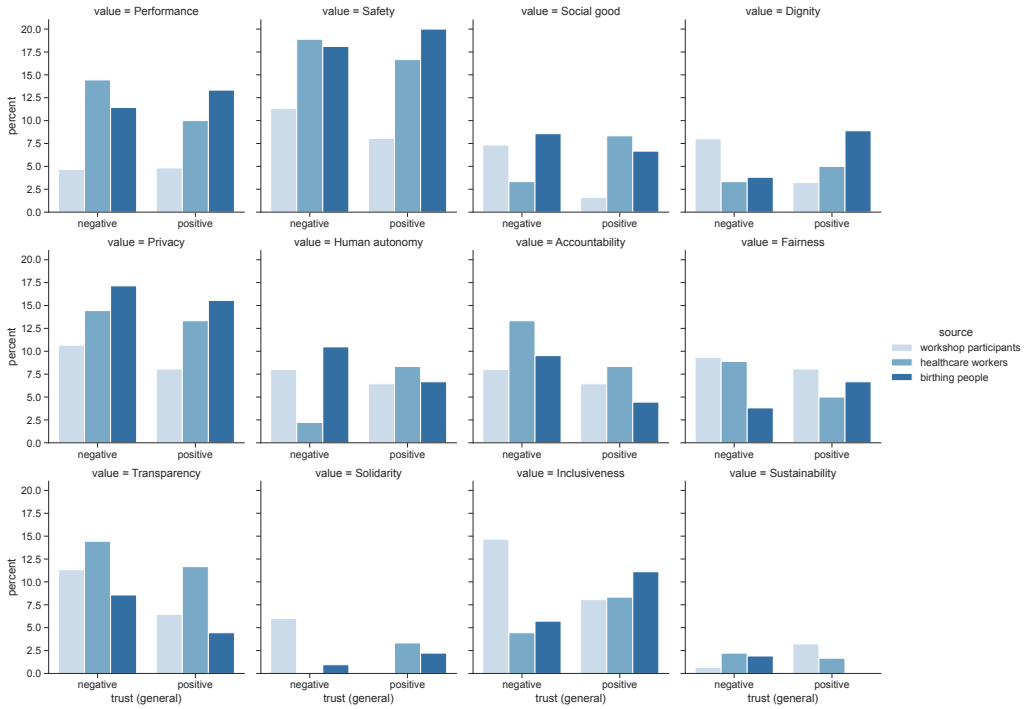

Fig. 7. Selection of values broken apart by cohort and **general** trust. *Positive* indicates scores > 3.

Alt Text: A set of bar plots showing the selection of values, with responses broken apart by participant cohort and generalized trust.

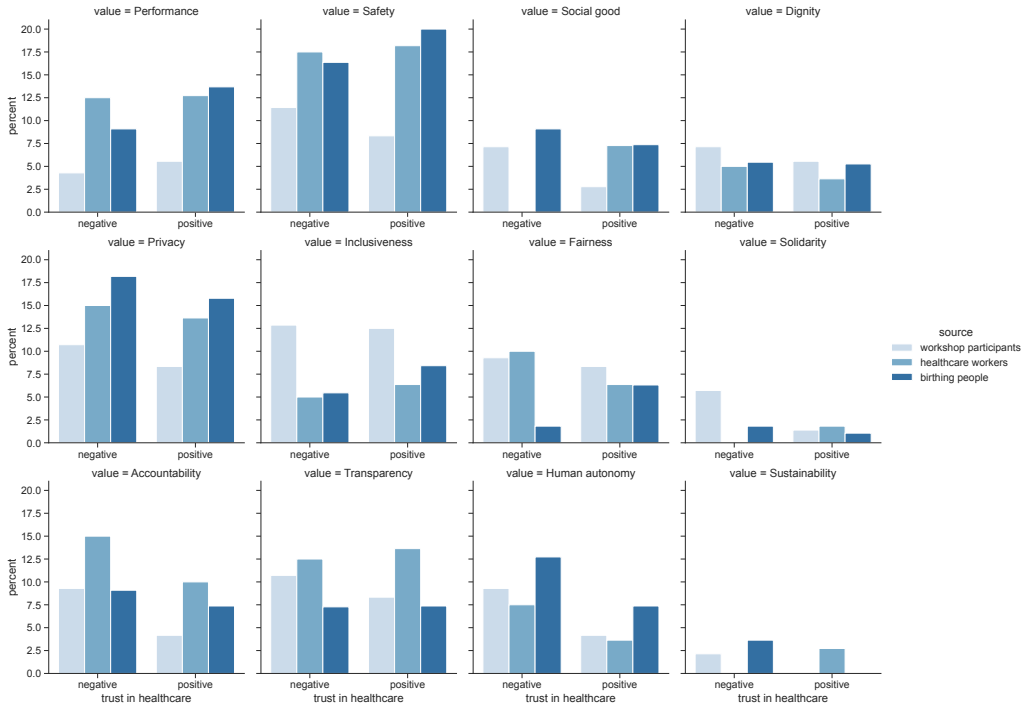

Fig. 8. Selection of values broken apart by cohort and trust in **healthcare providers**. *Positive* indicates scores > 3.

Alt Text: A set of bar plots showing the selection of values, with responses broken apart by participant cohort and trust in healthcare providers.

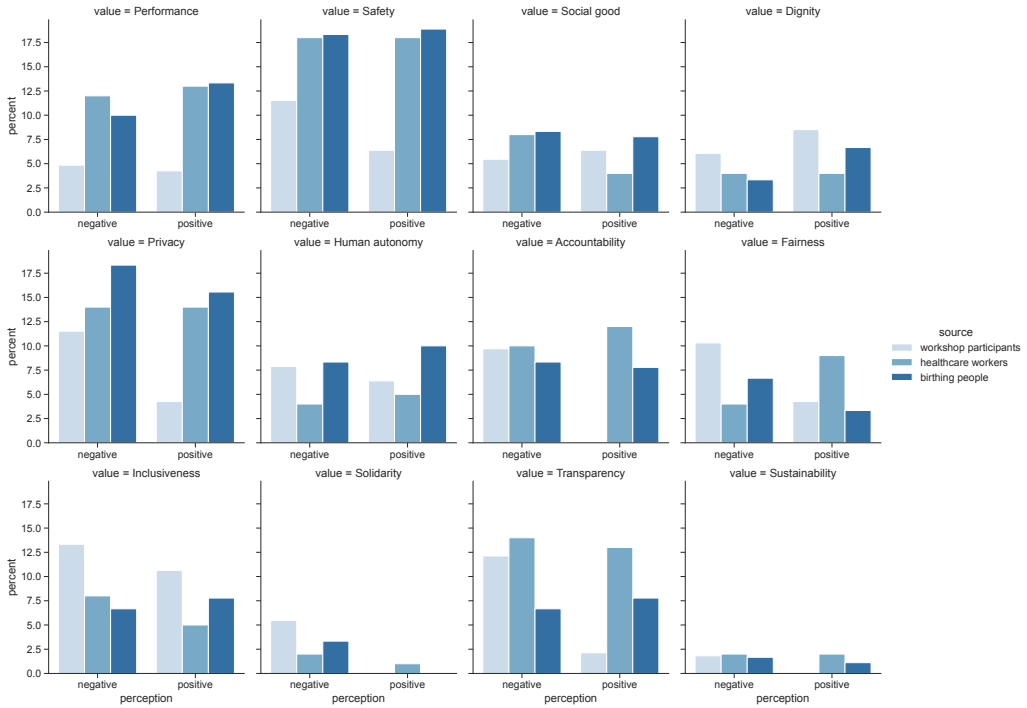

Fig. 9. Selection of values broken apart by cohort and general perceptions of AI. *Positive* indicates scores > 3.

Alt Text: A set of bar plots showing the selection of values, with responses broken apart by participant cohort and general perceptions of AI.

B CHATBOT DEMONSTRATION

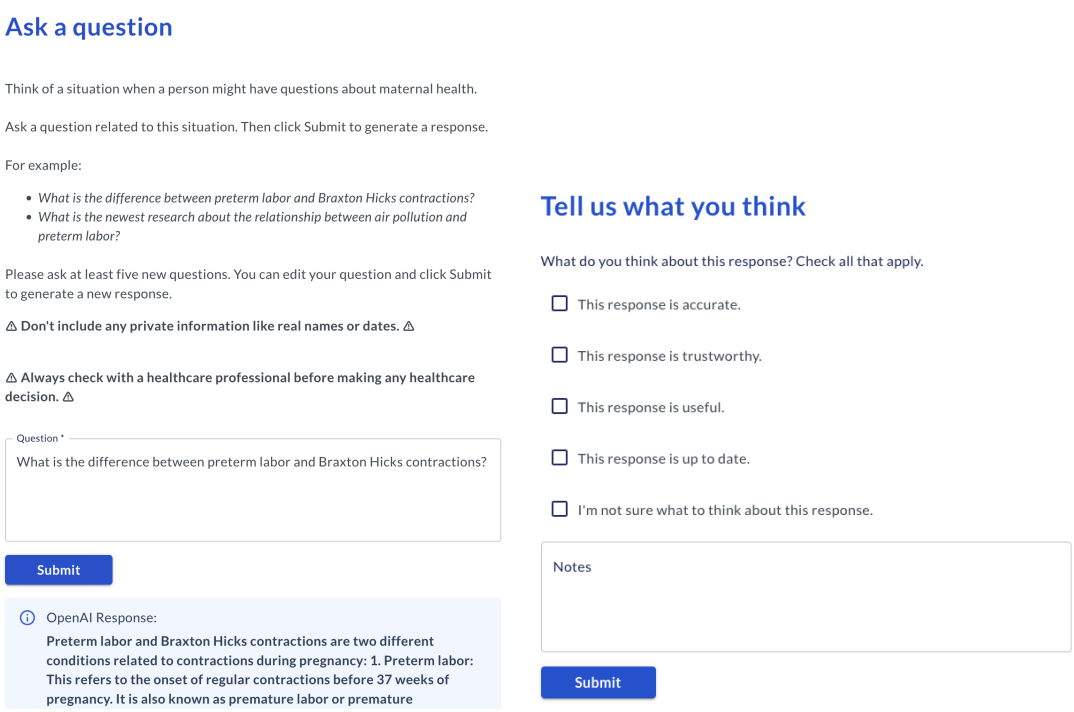

Fig. 10. Screenshots of our chatbot demo. Responses were generated using GPT-3.5 from OpenAI. Users could ask multiple questions about maternal health. The workshop demonstration additionally included a free-text entry box for participants to add descriptions of situations in which they might use the chatbot.

C FULL SURVEY QUESTIONS

C.1 Consent

Thank you for participating in this study about the use of AI for maternal healthcare. This study is being led by Maria Antoniak and Carla S. Alvarado. The purpose of this research is to gather perceptions about AI tools from a diverse group of people. This information will be used to create a set of guiding principles for the use of AI for maternal health. We deeply value your lived experience, and we want to include as many voices as possible in the development of these guiding principles.

Participation. Your participation in this study is voluntary and anonymous, and you may refuse to participate before the exercise begins, or discontinue at any time with no penalty.

In this study, you will complete a survey describing your experiences with maternal healthcare and your perceptions of AI tools.

Risks & Benefits. This study asks about your experiences in maternal health. Any questions related to this experience may make you uncomfortable or bring up feelings of stress. You are always free to decline to answer any question or stop your participation at any time.

You will be paid for your participation in this study according to the rates specific on Prolific. We hope to learn about the various perspectives in maternal health and represent these perspectives in the guidelines we develop, and we hope that these guidelines will shape future AI research for maternal health in ways that are useful, ethical, and appropriate for the community.

Privacy. We do not collect names or other personally identifiable information. We do collect your Prolific ID, but we do not have any way to link this with your identity. We will not release your Prolific ID to anyone else.

De-identified data from this survey may be shared with the research community at large to advance science and health. We will remove or code any personal information that could identify you before files are shared with other researchers to ensure that, by current scientific standards and known methods, no one will be able to identify you from the information we share. Responses will be stored securely by the Allen Institute for AI with limited access controls to limit exposure to those with a need to know.

Questions. If you have questions about this study, please reach out to Maria Antoniak or Carla S. Alvarado, or send us a message us on Prolific.

C.2 Screening Questions

(1) What is your Prolific ID?
(2) Do you consent to participate in this study?
(3) **[Birthing People Only]** Have you given birth in the last five years (2018-2023)?
(4) **[Healthcare Workers Only]** Have you ever worked in a healthcare profession?

C.3 Trust

(1) Generally speaking, would you say that most **people** can be trusted, or that you need to be very careful in dealing with people? *[Likert scale from (1) need to be very careful to (5) most can be trusted]*
(2) Generally speaking, would you say that most **healthcare providers** can be trusted, or that you need to be very careful in dealing with healthcare providers? *[Likert scale from (1) need to be very careful to (5) most can be trusted]*

C.4 Familiarity and perceptions of AI

In the following sections, we'll be asking about your familiarity and perceptions of AI systems. Some examples of AI systems include:

- A chatbot used by people to answer general questions
- An AI system used by a medical clinic to predict whether a patient has a disease
- An AI system used by a bank to predict whether an applicant will repay a loan
- An AI system used by a marketing company to match ads to viewers
- An AI system used by a streaming company to recommend movies to users

(1) How familiar are you with NLP, machine learning, and/or AI? *[Not at all familiar, Heard of it but never sought more information, Read about it, Occasionally used tools relying on these methods, Extensively used tools relying on these methods, Trained or finetuned my own models]*
(2) Have you ever used an AI chatbot like ChatGPT? *[Likert scale from (1) never to (5) all the time]*
(3) AI can eliminate a lot of tedious work for people. *[Likert scale from (1) strongly disagree to (5) strongly agree]*

(4) The overuse of AI may be harmful and damaging to humans. *[Likert scale from (1) strongly disagree to (5) strongly agree]*
(5) Life will be easier and faster with AI. *[Likert scale from (1) strongly disagree to (5) strongly agree]*
(6) AI turns people into just another number. *[Likert scale from (1) strongly disagree to (5) strongly agree]*

C.5 Values

These definitions are drawn directly from Jakesch et al. [40].

Fairness. A fair NLP system treats all people equally. Developers of fair NLP systems ensure, as far as possible, that the system does not reinforce biases or stereotypes. A fair system works equally well for everyone independent of their race, gender, sexual orientation, and ability.

Privacy. An NLP system that respects people's privacy implements strong privacy safeguards. Developers of privacy-preserving NLP systems minimize, as far as possible, the collection of sensitive data and ensure that the NLP system provides notice and asks for consent.

Sustainability. A sustainable NLP system preserves the environmental quality of current and future generations. Developers of sustainable NLP systems minimize, as far as possible, electricity use and reduce waste.

Inclusiveness. Inclusive NLP systems empower everyone and engage all people. Developers of inclusive AI systems consider, as far as possible, the needs of people who might otherwise be excluded or marginalized.

Safety. A safe NLP system performs reliably and safely. Developers of safe NLP systems implement strong safety measures. They anticipate and mitigate, as far as possible, physical, emotional, and psychological harms that the system might cause.

Social good. An NLP system that promotes social good supports, as far as possible, human well-being and flourishing, peace and happiness, and the creation of socio-economic opportunities.

Dignity. An NLP system that respects human dignity upholds the inherent worth of every individual. In addition to respective legislation, developers ensure that the system respects human rights and does not diminish human dignity.

Performance. A high-performing NLP system consistently produces good predictions, inferences or answers. Developers of high-performing NLP systems ensure, as far as possible, that the system's results are useful, accurate and produced with minimal delay.

Accountability. An accountable NLP system has clear attributions of responsibilities and liability. Developers and operators of accountable AI systems are, as far as possible, held responsible for their impacts. An accountable system also implements mechanisms for appeal and recourse.

Transparency. A transparent NLP system produces decisions that people can understand. Developers of transparent AI systems ensure, as far as possible, that users can get insight into why and how a system made a decision or inference.

Human autonomy. An NLP system that respects people's autonomy avoids reducing their agency. Developers of autonomy-preserving AI systems ensure, as far as possible, that the system provides choices to people and preserves or increases their control over their lives.

-

Solidarity. A solidary NLP system does not increase inequality and leaves no one behind. Developers of solidary AI systems ensure, as far as possible, that the prosperity as well as the burdens created by NLP are shared by all.

(1) Please select any five values from the list below that you think are most important for NLP systems for maternal health. *[Fairness, Privacy, Sustainability, Inclusiveness, Safety, Social good, Dignity, Performance, Accountability, Transparency, Human autonomy, Solidarity]*
(2) (Optional) Was anything missing from the list of values above? If so, describe here.

C.6 Information Seeking Behavior

(1) When I have questions about maternal health, I often turn to this resource to assist me.
[Likert scale from (1) strongly disagree to (5) strongly agree for each of the following options]
- Peer-reviewed scientific publications (for example: Journal of the American Medical Association)
- Online communities and forums (for example: Facebook groups, Reddit groups)
- Social media (for example: Instagram, TikTok, YouTube)
- Government resources (for example: FDA, CDC websites and announcements)
- News articles (for example: CNN, WSJ)
- Textbooks, books
- Expert-written websites (for example: WebMD, Mayo Clinic)
- Curated medical resources (for example: UpToDate)

(2) (Optional) Outside of the resources listed above, are there are other resources you turn to for help in answering questions about maternal healthcare?
(3) When I have questions about maternal health, I often turn to this person to assist me.
[Likert scale from (1) strongly disagree to (5) strongly agree for each of the following options]
- OB/GYN
- Nurse
- Midwife
- Doula
- Community health worker
- Healthcare reseracher
- Government workers
- Friends & family
- Colleagues

(4) (Optional) Outside of the people listed above, are there are other people you turn to for help in answering questions about maternal healthcare?

C.7 Effects of NLP/AI

(1) This person would benefit from an NLP/AI chatbot in their work.
[Likert scale from (1) strongly disagree to (5) strongly agree for each of the following options]
- OB/GYN
- Nurse
- Midwife

- Doula
- Community health worker
- Healthcare reseracher
- Government workers
- Friends & family
- Colleagues

(2) **[Workshop Only]** What impact do you expect an NLP chatbot would have on your work? *[Likert scale from (1) mostly harms to (5) mostly benefits]*

(3) **[Birthing People Only]** I would want my maternal healthcare provider to use AI systems. *[Likert scale from (1) strongly disagree to (5) strongly agree]*

(4) **[Birthing People Only]** I would want my maternal healthcare provider to tell me if they use AI systems. *[Likert scale from (1) strongly disagree to (5) strongly agree]*

(5) **[Birthing People Only]** I would want my maternal healthcare provider to tell me if they use AI systems. *[Likert scale from (1) strongly disagree to (5) strongly agree]*

(6) **[Birthing People Only]** How often do/did you consult internet resources for pregnancy-related questions during your pregnancy? *[Likert scale from (1) never to (5) frequently]*

(7) **[Birthing People Only]** If so, what were these pregnancy-related questions about? (check all that apply) *[medical worries/concerns during pregnancy, mental health worries/concerns during pregnancy, the pregnancy process in general, labor and delivery, birth control option after delivery, the postpartum process, taking care of a newborn, medical worries/concerns after pregnancy, mental health worries/concerns after pregnancy, Other (write in)]*

C.8 Demographics

All of these questions were optional.

(1) Which gender(s) do you identify with? *[woman, man, non-binary, prefer not to disclose, other (write in)]*

(2) What is your age range? *[<18, 18-24, 25-34, 35-44, 45-54, 55-64, 65-74, 74+]*

(3) Describe your race/ethnicity (check all that apply) *[African-American/Black, Middle Eastern/North African, Native American/Alaska Native/First Nations, Pacific Islander, Hispanic/Latino/a/x, East Asian (including Chinese, Japanese, Korean, Mongolian, Tibetan, and Taiwanese), South Asian (including Bangladeshi, Bhutanese, Indian, Nepali, Pakistani, and Sri Lankan), Southeast Asian (including Burmese, Cambodian, Filipino, Hmong, Indonesian, Laotian, Malaysian, Mien, Singaporean, Thai, and Vietnamese), White, Other (write in)]*

(4) Where do you live in the U.S.? *[Northeast, Midwest, South, West, I do not live in the U.S.]*

(5) **[Workshop Only]** What is your education level? *[Trade school (for example: AA, AS), College (for example: BA, BS), Master's degree (for example: MS, MA, MPH), Medical degree (for example: MD, DO), Other professional degree (for example: JD), PhD]*

(6) **[Workshop Only]** Describe your professional background. Check every profession that has **ever** applied to you. *[Community health worker/Promotoras, Community-based organization (non-profit), Doula, Certified Midwife, Certified Nurse Midwife, L&D Nurse, OB/GYN, Other perinatal health care provider, Other perinatal support (non-health care) provider, Non-perinatal health care provider, Health care management/administration, Health care services researcher, Population Health/Public Health researcher, Local/State public health entity, Federal Government*

-

employee in public health (for example: CDC, HHS, CMS and other related units)], AI / machine learning / NLP researcher or engineer, Other (write in)

(7) **[Workshop Only]** Have you ever sought maternal or reproductive support from a healthcare provider (for example: pregnancy support, contraception support)? *[yes, no, unsure, not applicable]*

(8) **[Prolific Only]** Do you or have you ever worked in healthcare (in a hospital or clinic, as a community health worker, as a public health researcher, or in any other capacity)? *[yes, no]*

(9) **[Prolific Only]** Do you or have you ever worked in maternal/perinatal healthcare (as a nurse, midwife, researcher, or in any other capacity)? *[yes, no]*

(10) Feel free to share any additional comments or feedback about this survey here. If anything was confusing or unclear, we'd love to know so that we can improve this survey in the future.