Powered by AI: Examining How AI Descriptions Influence Perceptions of Fertility Tracking Applications

MAYARA COSTA FIGUEIREDO, Universidade Federal do Pará, Brazil
ELIZABETH ANKRAH, University of California, Irvine, USA
JACQUELYN E. POWELL, University of California, Irvine, USA
DANIEL A. EPSTEIN, University of California, Irvine, USA
YUNAN CHEN, University of California, Irvine, USA

Recently, there has been a proliferation of personal health applications describing to use Artificial Intelligence (AI) to assist health consumers in making health decisions based on their data and algorithmic outputs. However, it is still unclear how such descriptions influence individuals' perceptions of such apps and their recommendations. We therefore investigate how current AI descriptions influence individuals' attitudes towards algorithmic recommendations in fertility self-tracking through a simulated study using three versions of a fertility app. We found that participants preferred AI descriptions with explanation, which they perceived as more accurate and trustworthy. Nevertheless, they were unwilling to rely on these apps for high-stakes goals because of the potential consequences of a failure. We then discuss the importance of health goals for AI acceptance, how literacy and assumptions influence perceptions of AI descriptions and explanations, and the limitations of transparency in the context of algorithmic decision-making for personal health.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; • **Applied computing** → **Consumer health**.

Additional Key Words and Phrases: Consumer health technologies, Artificial intelligence perceptions, Fertility self-tracking

ACM Reference Format:
Mayara Costa Figueiredo, Elizabeth Ankrah, Jacquelyn E. Powell, Daniel A. Epstein, and Yunan Chen. 2023. Powered by AI: Examining How AI Descriptions Influence Perceptions of Fertility Tracking Applications. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 4, Article 154 (December 2023), 24 pages. https://doi.org/10.1145/3631414

1 INTRODUCTION

With the advances in artificial intelligence (AI) and the ever-increasing availability of data, there has been a proliferation of consumer health applications (apps for short) combined with AI. Covering a wide spectrum of lifestyle choices and health conditions [64, 82], these apps intend to assist people in making health decisions based on their data. For example, AI-based symptom-checker apps use AI to assess users' symptoms and provide potential diagnostic results [89], AI-based chatbots are increasingly common for supporting mental health [46, 63] and AI-infused fitness apps gained significant traction in the COVID-19 pandemic [48]. While these systems have great potential to empower individuals to take care of their health, they often offer little transparency on

Authors' addresses: Mayara Costa Figueiredo, mcfigueiredo@ufpa.br, Universidade Federal do Pará, Rua Augusto Correa, 1, Belém, Pará, Brazil; Elizabeth Ankrah, University of California, Irvine, Irvine, California, USA, eankrah@uci.edu; Jacquelyn E. Powell, University of California, Irvine, Irvine, California, USA, jacqueep@uci.edu; Daniel A. Epstein, University of California, Irvine, Irvine, California, USA, epstein@ics.uci.edu; Yunan Chen, University of California, Irvine, Irvine, California, USA, yunanc@ics.uci.edu.

2474-9567/2023/12-ART154 $15.00
https://doi.org/10.1145/3631414

how their algorithms work [37, 82], providing vague descriptions or behaving as black-boxes (i.e., offering basic justification for the system and occasionally a description about the data it uses, but without disclosing how it generates recommendations [13, 71, 82]). Previous studies have also reported that AI-related terminology can be used as a selling argument [49] and to signal increased technological precision [15]. Terms such as 'artificial intelligence,' 'smart algorithms,' or 'machine learning' are often used interchangeably to convey a high level of predictive accuracy [82], or even attract investments and imply technological innovation [76, 85]. It is still unclear how this context influences individuals' perceptions, understanding, and trust in AI-infused health apps, especially when these technologies are intended for a public that may not have the domain (e.g., health) or technology (e.g., AI) expertise to assess algorithmic outputs when making important personal health decisions. Understanding this perception is particularly important in light of conversations about personally tracked health data that is shared publicly online being used to train AI models [35, 68] and recent discussions around privacy and intimate data [77].

This study investigates how AI-related descriptions influence individuals' perceptions of personal health technologies. To do this, we focus on fertility self-tracking, a health domain that has been increasingly influenced by AI. Fertility wearables and apps are increasingly leveraging AI and machine learning to provide users predictions for period dates, ovulation, and fertile window based on self-tracked data. Previous studies have reported that fertility apps do not commonly explain how these predictions are generated or how collected data are used [28, 32, 43, 65, 77], although apps' descriptions often emphasize their accuracy and support in enabling control over the body [19, 20]. Furthermore, research in algorithmic perceptions emphasize that the type of task is a strong factor influencing individuals' attitudes towards algorithmic decision-making [4, 49, 51]. Fertility includes multiple goals that influence tracking in different ways and that have different potential consequences, which can be used to explore different tasks within the same domain.

We therefore investigate the research question: **how might AI descriptions influence individuals' perceptions of fertility tracking apps in the context of different fertility goals?** To do so, we created three versions of a simulated fertility app, Kaya, varying the way the app uses (or not) AI-related terms to refer to its features (i.e., AI descriptions). 298 US participants used one random version of Kaya and answered a survey on their perceptions about the app. Overall, participants expressed positive reactions to Kaya, expected that versions mentioning AI (AI versions for short) to be more accurate, and were more willing to download and use them. Users trusted AI versions more for avoiding conception, but qualitative responses reveal that although participants expect apps using AI terms to be accurate and generally prefer them, they were unwilling to rely solely on these apps for more high-stakes goals (i.e., avoiding conception) because of the consequences if the app failed. Responses also show participants assume how AI works, intertwining these assumptions with their beliefs of technology, which together influence their perceptions of AI descriptions in different ways.

This study contributes to AI perceptions and personal informatics (PI) literature by investigating the influence of AI descriptions on individuals' personal health and suggesting how to improve their design in light of people's perceptions. Our study also contributes understanding of how different tasks within the same health domain influence these perceptions. We also contribute a discussion on how AI assumptions and individuals' literacy interact with AI descriptions, both positively and negatively, and what this interaction means for the design of AI-infused personal informatics (PI) systems to inspire appropriate trust.

2 RELATED WORK

To analyze individuals' perceptions of AI descriptions in consumer health technologies, this study builds on previous research on perceptions of algorithmic recommendations, PI-influenced decision-making, and fertility self-tracking.

2.1 Perceptions of Algorithmic Recommendations

Extensive research has focused on how people trust algorithmic decisions in comparison to human made ones, particularly for managerial and work-related contexts [11, 49, 51]. Such studies often suggest that people judge algorithmic decisions as inferior or similar to human decisions, being particularly more negative for tasks that are subjective [11], demand human capabilities [51] or require attention to individuals' unique characteristics [58]. These studies emphasize that people's attitudes towards or reliance on algorithmic recommendations are largely influenced by the type of task the system is intended to support or automate [4, 49, 51].

Any system that processes data (e.g., averaging values, showing trends in data points) and generates recommendations can be used to support human decision-making. In this paper we focus on AI-based systems because of the growth in AI advertising (particularly in health apps [5, 82]) and because AI introduces additional challenges (e.g., inconsistent and unpredictable behaviors, information hiding, accuracy and error communication [2, 26, 47, 72]) to the health domain. In fact, Longoni et al. [58] highlight healthcare has intrinsic characteristics that influence resistance to medical AI: most health domains are unfamiliar to most people, abound with uncertainty, and often involve life-threatening consequences.

Besides resistance, trust has been suggested as an important determinant for adoption of systems to support decision-making [51]. For example, Lee and Rich [53] investigated the role of mistrust in human decisions on individuals' perceptions of AI systems. Focusing on US Black and African American individuals (who have higher medical mistrust due to anti-Black racism they face across human-led institutions), they found that participants with high cultural mistrust perceive healthcare AI as equally untrustworthy and unfair as human medical providers, and suggest providing more information about AI algorithms, development, and usage to improve people's trust.

In contrast with these studies, Araujo et al. [4] found that in high impact health tasks, decisions made by a hypothetical AI system were, in general, associated with more positive attitudes (less risky, fairer, and more useful) than decisions made by human experts. Similarly, Kapania et al. [45] used the concept of AI authority (i.e., *"the legitimized power of AI to influence human actions, without requiring adequate evidence about the capabilities of the given system"*) to explain the overall high acceptance of AI decisions and intentions to use in both low and high-stakes tasks in India. Their study highlights the importance of external narratives and societal influences on the ways people perceive and accept AI.

All these studies use human decision making as a benchmark to evaluate people's perceptions of described AI systems. In these cases, people may be more critical of the systems because they have the expert decision as a counterpoint. Furthermore, other studies report that individuals' attitudes differ when human decision making is not used as a comparison point or when individuals compare algorithmic recommendations to their own estimates [57]. Inspired by these studies, we investigate trust and willingness to download and use in a context where individuals often do not have expert support.

2.2 Algorithmic Decision-making in Personal Life

With the extensive commercialization and use of mobile and sensor devices, there has been a proliferation of personal health systems, typically called personal informatics (PI) or self-tracking systems, designed to help individuals in collecting and reflecting on their personal data so they can make informed personal health decisions [29, 54]. Besides collecting and storing personal data, PI tools often summarize and process these data to produce recommendations and support users' interpretation and action [7].

In general, PI tools offer little transparency as to how their algorithms work, how reliable the results are, and what personal data are collected and used in making those algorithmic determinations [7, 88]. Previous PI studies describe that people's perception of accuracy may be disconnected from their systems' capabilities. For example, Yang et al. [88] have noted that *"the black-box nature"* of PI algorithms can *"inhibit users from understanding"*

how they work, which can be negative especially for non-expert end-users when they heavily rely on automated systems [6]. Hollis et al. [39] reported that users might defer to an algorithm's classification of their emotional experience over their personal judgment of that experience. Similarly, Warshaw et al. [86] describe people can *"have greater confidence"* in an algorithm than in their ability to describe their own personality. Research in algorithmic decision-making and PI have also reported that perceived accuracy influences adoption, use, and trust [24, 49, 51, 53, 66, 88]. These expectations of accuracy suggest that users may place inappropriately high levels of trust in *"black-boxed"* algorithms for personal use [80].

These studies have suggested that lack of transparency and expertise to challenge the data or recommendations have impacted user trust, adoption, and use of PI apps which do not explicitly use AI. Incorporating AI can exacerbate these challenges and add specific challenges related to AI [2, 26, 47, 72]. While research has investigated how to use AI and machine learning to improve PI tools [24, 34, 55, 70, 84], there has been a fast growth in AI use in commercial consumer health technologies covering a broad spectrum of lifestyle choices and health conditions [5, 82]. AI terms that convey predictive accuracy and technology innovation [76, 82, 85] add to this context and may further influence individuals' perceptions of AI-infused PI tools. For these reasons, we also investigate individuals' expectations of accuracy.

2.3 Fertility Self-tracking

The fertility of individuals who menstruate (fertility for short) is complex [17], entangled with stigma and social taboos [1, 23, 44], and many people have only low to intermediate knowledge about it [9, 59]. Fertility self-tracking apps allow end-users to collect varied health indicators potentially related to their cycles (e.g., period dates and other physical and emotional data), process these data, and provide feedback for users to support their health decisions [17]. Recently, fertility self-tracking has drawn significant attention in the consumer technology market. Fertility apps, which were downloaded around 200 million times worldwide back in 2016 [27], constitute a major part of the so-called "Femtech"[1] industry, a market that has been estimated to reach a value of $50 billion by 2025 [33]. In this study, we focus on individuals who menstruate because most indicators tracked in fertility apps are specific bodily phenomena (e.g., menstruation, cervical mucus) which are more often tracked and reviewed only by themselves.

People track fertility for multiple goals, including, to conceive, avoid pregnancy, and assess their body status [30]. A major part of how fertility apps support these goals involve predictions for periods, ovulation, and fertile window. People use these predictions differently to make appropriate decisions towards their goals, e.g., from refilling medications for pre-menstrual symptoms, to having intercourse for conceiving. So, fertility predictions are related to decisions that can contribute to consequences that are emotionally-loaded (e.g., increased hopes and severe frustrations when trying to conceive [16]) and life-changing (e.g., pregnancy).

Many fertility apps are currently advertised as using 'artificial intelligence,' 'smart algorithms,' or 'machine learning' to convey predictive accuracy [15, 82]. Nevertheless, most apps do not explain their algorithms or how they generate predictions [28, 32, 43, 65, 75], often providing ambiguous descriptions about their AI use [31]. It is also not clear what is done with the varied health indicators collected via these apps and what data among these are used to create fertility predictions [20, 62, 77]. This scenario makes fertility a high-stakes context, in which AI descriptions and app output may influence critical decisions that can lead to important personal consequences. In this work, we focus on three major goals within the same health domain: period tracking, trying to conceive, and avoiding conception. Unlike tasks within different health contexts (fitness recommendations vs. treatment decisions [4]), these goals include similar data collection and analysis, but the health decisions and their consequences are different (e.g., having a period without expecting vs. an unplanned pregnancy).

[1]Femtech stands for Female Technology and encompasses a variety of technological products focused on the health of individuals who menstruate

3 KAYA OVERVIEW

We developed Kaya to explore how AI descriptions influence individuals' perceptions (i.e., expectations of accuracy, trust, and willingness to download and use) of fertility self-tracking apps in the context of different goals. Like similar PI studies [7], Kaya's design was inspired by the most popular commercial fertility apps [19, 30]. We designed Kaya as a simulated app so participants could experience using the main features, including data input and feedback, before answering questions on their perceptions of such an app [49]. In the simulation, participants were led through the process of reading the app store page, downloading the app (a simulated download, nothing was installed in participants' devices), inputting fertility-related data, and visualizing a calendar and a graph. Overall, it took about 6 minutes to simulate the use of Kaya by passing through its main screens (Figure 1): (a) the app-store page, (b) the calendar screen, (c) the input screen, and (d) the graph screen. These screens model common screens of commercial fertility apps and contain the main features such apps offer [19]. A dialog with instructions precedes each screen.

The app store page (Figure 1a) describes Kaya's main features. A dialog asks participants to read the page and download the app (reminding them that nothing is downloaded or installed via the simulation).

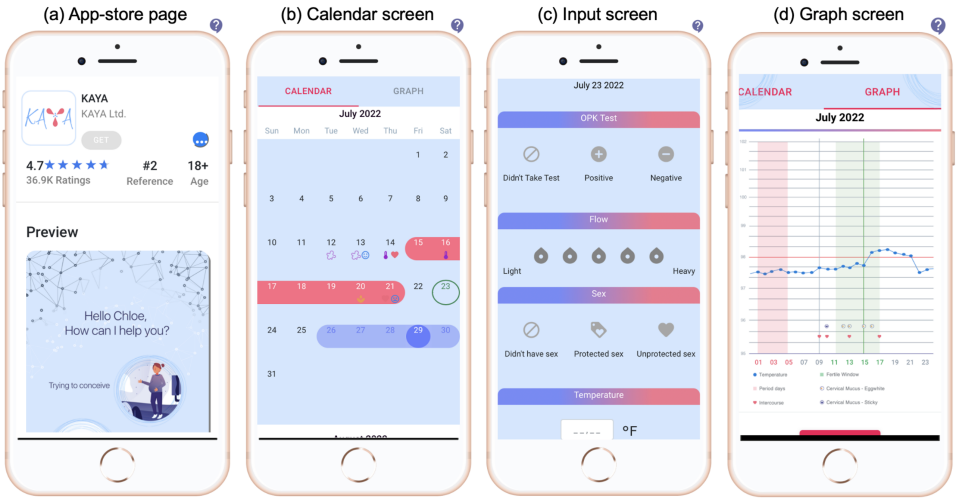

Fig. 1. Kaya main screens: language, graphs, icons, and features replicate currently available fertility apps.

The first Kaya screen is the calendar (Figure 1b), which is loaded with fictitious data, always showing the previous period before and the fertile window after the current day, following the generic 28-day fertility cycle [79]. Icons appear on some days to exemplify how Kaya displays previously recorded data in the calendar. The instruction dialog explains all the symbols and colors and instructs participants to click in the current day to access the input screen (Figure 1c).

The input screen is interactive, allowing participants to input data for all available health indicators (chosen based on the main indicators offered by fertility apps [19]), including period dates, intercourse, temperature, ovulation predictor kit's results (OPKs), and pregnancy test results (Figure 1c). Participants are instructed to log the data described in a scenario (i.e., fictitious data instead of their personal information) so they can experience tracking using Kaya. When they save the data, the simulation displays a loading dialog with a message briefly describing what Kaya is doing with the data and asks them to access the graph. The graph screen (Figure 1d)

shows an image of a temperature graph, and participants are asked to analyze the image and access the survey. Next, we describe how Kaya was used in the study, including how it varied across versions.

4 METHODS

To answer our research question, we used a convergent mixed methods design [22] in which participants used one version of Kaya and answered a survey with both quantitative and qualitative questions. The following subsections describe the study design, the survey, participants' characteristics, and how quantitative and qualitative data were analyzed and integrated.

4.1 Study Design

The study started with a screen containing its description, information sheet, eligibility criteria, and consent. Once consent was obtained, participants were informed that they would be reviewing a prototype of an app for fertility tracking, but that the app was not fully functional or commercially available. They received access to one of three versions of Kaya:

- Base version, with descriptions without AI terms and without explanations for predictions,
- AI Keywords version, which adds AI terms to the Base version without adding explanation, and
- AI Explanation version, which adds explanation to the previous one by informing the data the app uses and how it starts predicting.

The idea behind these versions was to (i) mimic the most common descriptions of commercial fertility apps with our Base and AI Keywords versions, and (ii) compare them with a third version that contains an explanation, representing an improvement towards transparency. In summary, the Base version simulates current apps that do not mention AI and do not explain their algorithms. The AI Keywords version adds AI terms beyond the Base (Figure 2) to mimic many current apps that advertise using AI but do not explain how it works or provide explanations for predictions [28, 32, 43, 65]. Then, inspired by Newn et al.'s [66] finding that the presence of explanations influences individuals' acceptance of personal sensing technology, we designed an AI Explanation version to understand how participants perceive and value AI explanations in comparison with the other two. This version mentions the app uses AI, but it also explains which data it uses, how predictions start, and how uncertainty is displayed (which is not commonly observed in commercial fertility apps [75]) (Figures 2 and 3). Similar to Newn et al. [66], we did not intend to test the full design space of explanation styles (e.g., where the explanation is presented in the app, text versus graphical explanations). We specifically sought to analyze how the presence of an explanation influences individuals' perceptions and acceptance of the app.

The three versions differed in six main ways: in (i) the logo, (ii) the app store page, (iii) the calendar page, (iv) the calendar instruction dialog, (v) the loading dialog that appears after logging data, and (vi) the instruction of the graph page. As an example, Figure 2 shows an excerpt of the app-store page to illustrate how descriptions varied across versions.

Figure 2 shows that the AI Keywords version only adds "smart algorithms", "machine learning" and "artificial intelligence" terms to the Base description. Langer et al. [49] discuss that terminology can be used strategically to influence laypeople's perceptions of systems. Inspired by this work, we chose these terms we observed in many fertility apps (and health apps in general [5, 31, 82]) so we can appeal to similar emotions and perceptions they may influence on consumers [49, 82]. We kept the same AI-related terms in the AI Explanation version to appeal to the same emotions and perceptions [49].

Other important differences are observed in the calendar screen (Figure 3 left) and the loading dialog (Figure 3 right). The calendars for the AI keywords and Base versions are the same. The difference resides in the instruction dialog that mentions that the AI keywords version uses AI. The calendar for the AI Explanation introduces uncertainty through the gradient color around the ovulation day and the dashed lines around the fertile window,

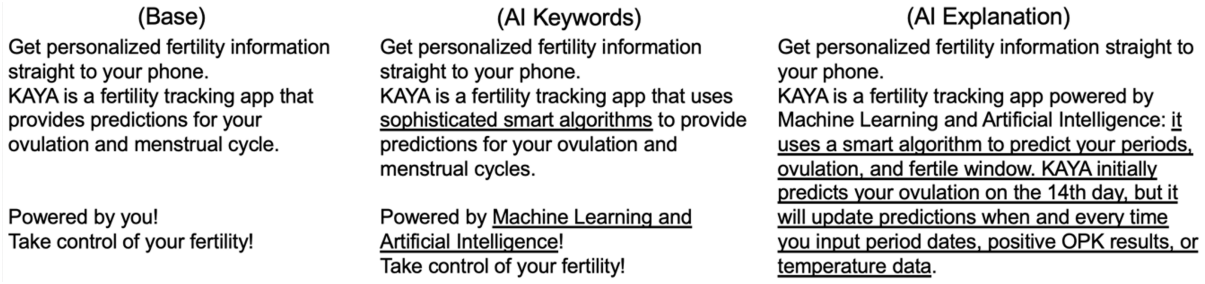

Fig. 2. The app store page for the Base and AI Keywords versions follow typical descriptions of commercial fertility apps. AI Keywords, as the name suggests, adds AI-related terms that have been increasingly used in AI-enabled commercial apps and media coverage. In contrast, the AI Explanation version adds which data it uses and briefly explains how Kaya starts generating predictions on the top of AI Keywords text. This version is intended to provide more information to users to analyze its impacts.

which is also described in the instruction dialog) to show potential errors in the future [75]. The loading dialog (Figure 3 left) appears after users input data and displays what Kaya does with them, from simply storing the data in the Base version to describing what data Kaya is using to improve predictions in the AI Explanation version.

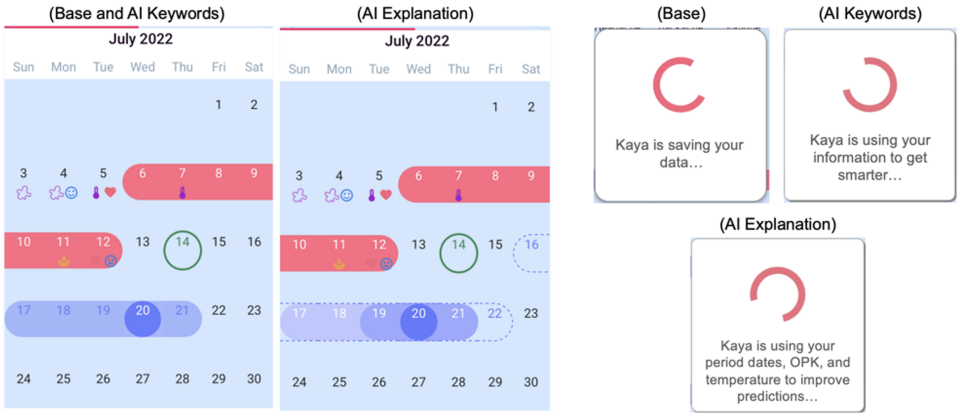

Fig. 3. In the Calendar screen (left) we varied how prediction accuracy was presented between versions, with the Base and AI Keywords versions indicating high probability for a single ovulation date and the AI Explanation version highlighting uncertainty over a range of days. The loading dialog (right) gives progressively more information on what Kaya does with input data from Base to AI Explanation.

Because of recent discussions related to intimate data and the recent overturn to the right of abortion in the US and its potential consequences [67, 83], we opted to mitigate the influence of privacy concerns regarding the collected data on participant's responses. We therefore described in all three versions that Kaya does not collect location data and all data inputted in Kaya is stored only locally in individuals' phones. This description was present in the app store page.

Participants were randomly assigned to one of the versions. We did not make participants walk through all three to avoid survey fatigue: the study took an average of 29 minutes to complete only through one version. After participants completed the simulation, they were asked to answer a survey.

4.2 Survey Design

Survey questions focused on individuals' perceptions of fertility tracking apps in the context of major fertility goals, i.e., period tracking, trying to conceive, and avoiding conception. We understand that there are more goals to track fertility, but we aimed to mirror the goals that most apps aim to support [19, 30]. We focused on the three following aspects to characterize overall perceptions of apps' descriptions. All Likert questions followed a 7-point scale with follow-up open-ended questions asking participants to explain their reasoning.

- *Expected accuracy*: we evaluate expected accuracy to gather insight into how useful participants might expect a version to be, how hopeful they might be about it, and how app descriptions support or reinforce those expectations. We asked participants how accurate they expected Kaya's predictions for period dates and for ovulation and fertile window to be.
- *Trust*: we investigate participants' self-reported trust prior to use focusing on each of the main three goals to evaluate not only how descriptions influence individuals' trust in the app, but also how these perceptions may change based on goals and the tasks that support them. We asked why participants trust or distrust Kaya predictions and how likely they would be to trust Kaya for each of the studied goals.
- *Willingness to download and use*: we analyze how the different descriptions influence participants' decisions whether or not to install and use an app for each goal. We asked participants how likely they would be to download Kaya, how they think Kaya makes predictions, how likely they would be to use Kaya for the three goals, and why.

After these questions, we presented to participants an image of the app store page of another version, assigned randomly, to investigate how they perceive different app descriptions and enable some within-subjects comparison. We asked questions comparing both apps' descriptions, replicating questions about download, use, and perceived accuracy. Finally, we included questions on general fertility knowledge, use of fertility apps and technology, and demographics, with open-ended fields, multiple selection (for ethnicity, gender, and sexual orientation), and optional responses for demographics [74]. Because we ran the study around the time the US Supreme Court overturned Roe v. Wade (right to abortion) [67, 83], we also asked participants how, if at all, this news impacted their perception of Kaya and other similar apps. The supplementary materials contain the list of questions used in the study. Figure 4 summarizes the complete study process.

We iteratively developed and tested the survey through revision among the authors and feedback from a pilot with our research group and colleagues. The study was classified as exempt by our institution's IRB since the survey methodology does not involve more than minimal risk to participants and no identifiable information was collected. We met with our IRB office prior to data collection to get advice on how to broach this sensitive topic with participants, particularly in light of abortion conversations. Our IRB ultimately concluded that our questions offered no more than minimal risk and did not run the risk of concerning self-disclosure, such as a participant disclosing to us that they had an abortion. We further did not observe any such disclosures in our participant responses.

4.3 Recruitment

We recruited participants on Prolific [69] in June 2022. We used Prolific screening tool to recruit people between 18 and 55 years old (most people reach menopause between 45 and 55 years old [12]), living in the US, fluent in English (since the simulation and survey are in English), and that self-describes their sex assigned at birth as female, prefer not to say, or both male and female. Before the study began, we also screened for people who have or had a menstrual cycle. We paid each participant $5 based on Prolific's suggestion for a 30-minute survey.

We recruited 298 participants and analyzed their responses. The supplementary materials show participants full self-identified demographics. In summary, the study's population comprised mostly highly educated (208 or 69.8% had at least an associate degree), urban (235, 78.86%), young (ages ranged from 18 to 55, mean=28.33, sd=7.37),

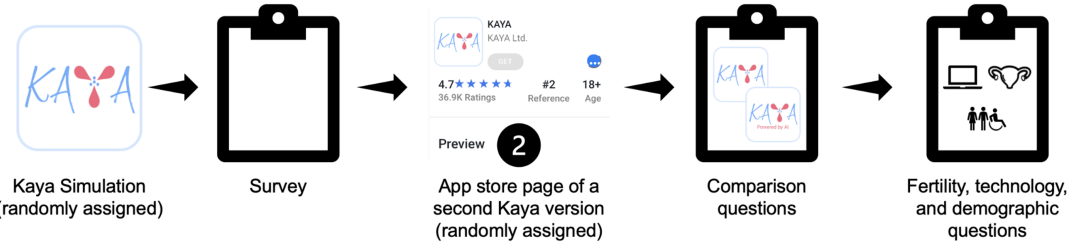

Fig. 4. Study process: participants first simulated the use of a random version of Kaya and answered survey questions on their opinions and attitudes towards the app. They were then presented the app store page of a second app assigned randomly and answered questions comparing both apps' descriptions, including replicated download, use, and accuracy questions. Finally, participants answered fertility, technology, and demographics questions.

heterosexual (204, 68.45%) or bisexual (55, 18.46%) women (283, 94.97%) who have or had periods. Although most participants were white or Caucasian, this proportion was lower than in the general US population (46.64% vs. 64.1%) while the proportion of Black or African American participants was higher in our sample (24.83% vs. 12%) [10]. The proportion of the other races or ethnicities was similar to US demographics [10]. However, it is important to note that the study's population did not include American Indian and Alaska Native participants and only one participant selected Native Hawaiian and Other Pacific Islander in combination with Hispanic or Latino, limiting the representation of these populations. Participants had higher educational attainment than the mean of the US population (32.1% vs. 77.9% of our population 25+ years old has at least a Bachelor's degree) [61] and the median household income (between $40,000 - $59,999) was below the median for the US ($67,521) [78].

Goals for using Kaya were split, following expected breakdowns [30]: 198 (66.44%) participants chose period tracking, 52 (17.45%) avoid conception, 43 (14.43%) try to conceive, and 5 (1.68%) other reasons, describing multiple goals or communicating with medical providers. This ratio emphasizes the experiences of people who would primarily use an app for period tracking. Participants considered themselves knowledgeable about fertility (a lot of knowledge: 95 participants, 31.88%; some knowledge: 165, 55.37%), most have regular cycles (195, 65.44%) and do not have children (225, 75.50%) nor tried to conceive (235, 78.86%). They are also experienced in tracking periods and using fertility apps: 236 (79.19%) have tracked their menstrual cycles and 186 (62.42%) have tried a period tracking app, with 90 (30.20%) using one at the time of the study. Most participants felt confident using fertility apps (median = Somewhat agree, 203 positive responses) and reported understanding how they calculate predictions (median = Agree, 234 positive responses). Experiences with AI were not widespread (43.29% reported they had not used a technology system that uses AI, machine learning, or data science, while 34.56% did not know), although most participants describe understanding (median = Somewhat agree, 243 positive responses) and trusting (median = Somewhat agree, 221 positive responses) such systems.

4.4 Analysis

Of the 298 participants, 98 were randomly assigned to the Base version, 91 to the AI Keywords, and 109 to the AI Explanation. We dropped the 5 participants who described a goal other than the main three from goal-related tests to simplify analysis.

To identify potential differences among groups, we run Kruskal–Wallis tests for ordinal and Chi-squared tests for categorical dependent variables from the demographics, fertility knowledge, and technology use questions, using both the app version and goals as independent variable. For trust questions, we ran Kruskal-Wallis tests using the app version as independent variable. For accuracy, download, and use questions we used ordinal regression models using participants' answers to both apps they compared in the survey, adding a variable to

consider the order effect. We treated the Likert ratings answers as ordinal and the fixed effects as categorical, examining app version (three levels), order of visualization (two levels), and interactions between them. For questions focusing on specific goals (i.e., trust and use), we run the tests on subsets of the data for each goal to exclude speculation on behalf of the participants whose primary goal was different than the one specified in the question. For all tests with significant results, we ran effect size statistics (epsilon-squared for both Kruskal-Wallis and ordinal regressions) and post-hoc tests (Dunn Test for Kruskal-Wallis and Estimated marginal means for ordinal regressions) to identify which groups differed. We report effect size values alongside their interpretation (i.e., small, medium, large) following values presented in [60]. All tests were performed using R.

We used the qualitative responses to interpret the quantitative findings. One researcher first read all the answers to get a sense of their content and discussed preliminary results with the other authors. Then the same researcher conducted a mix of inductive and deductive analysis using structural and initial coding in the first cycle and pattern coding in the second cycle coding [73] focusing on each of the three aspects related to individuals' perceptions. Examples of second-cycle codes and some of their subcodes include "Feelings from descriptions: Base empowers users, AI takes control from users, Base is warmer, AI is colder", "AI Assumptions: AI is less work, AI is more accurate, All apps use AI", "Transparency: Appreciates explanation, Too much explanation, AI is enough explanation", "Privacy: worries about apps, apps more important now", "Consequences: won't use for TTA, Kaya + other BC, TTC low-stakes, TTC sensitive"[2], "Accuracy: accurate because I provide data, need to test". Qualitative results were iteratively discussed among the authors during the analysis.

5 RESULTS

Participants had positive attitudes towards all three versions of Kaya. Overall, participants expected versions that mention AI to be more **accurate**, were more **willing to download and use** these versions (particularly the AI Explanation), and **trusted** them more for avoiding conception (both versions but particularly the AI Keywords). Despite this preference for AI versions for avoiding conception, qualitative responses show that participants were unwilling to use Kaya as their only birth control means because of the potential consequences in case of a failure. Qualitative responses further show different beliefs and assumptions participants had on how AI works and what personalization means based on the provided descriptions. We identify participants by a P followed by a number. We did not observe significant differences among groups based on demographics, fertility knowledge, and use and attitudes towards technology and fertility apps.

5.1 Expectations of Accuracy: "Accurate Information Provides Accurate Results" (P232)

Overall, participants expected Kaya' predictions to be accurate (Figure 5). Some qualitative responses indicate that participants believe apps would be accurate and their calculations will be correct if they diligently provide correct information, as P247 summarizes: *"I considered that if I provided Kaya with true information, then I have no doubt that the predictions will be correct."* Quantitative findings suggest that apps' descriptions may reinforce these expectations.

We found a small effect of both app version (E^2=0.038 for period dates and E^2=0.040 for ovulation and fertile window) and order (E^2=0.069 and E^2=0.079 respectively) in participants' expectations of accuracy. Participants expected that toward both predicting period dates and ovulation and fertile window, the AI Explanation (period dates: z=4.61, p<0.001, 95%CI 0.50-1.23 higher on a 7-point Likert scale; ovulation and fertile window: z=z=4.81, p<0.001, 95%CI 0.53-1.25 higher on a 7-point Likert scale) and AI Keywords (period dates: z=2.26, p=0.024, 95%CI 0.06-0.79 higher on a 7-point Likert scale; ovulation and fertile window: z=3.034, p=0.0024, 95%CI 0.20-0.93 higher on a 7-point Likert scale) would be more accurate than the Base version. The post-hoc test showed that participants expected the AI Explanation to be the most accurate for predicting period dates—it was rated more

[2]BC = Birth Control, TTC = Trying to Conceive, TTA = Trying to Avoid Conception

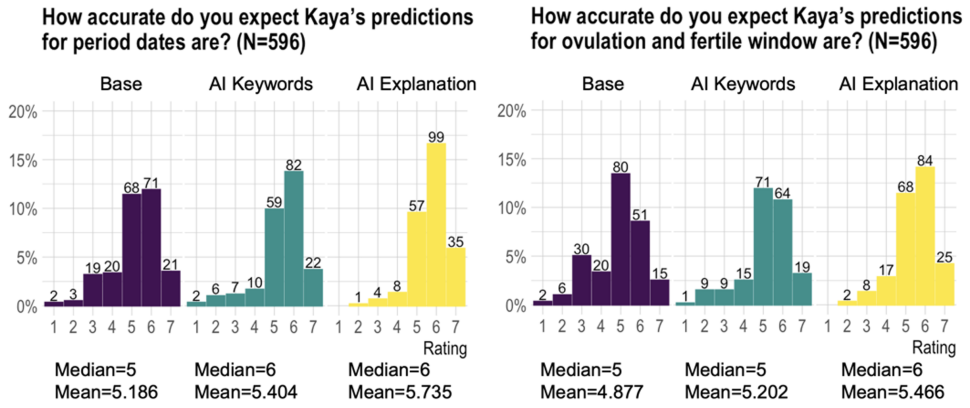

Fig. 5. Participants' expectations of accuracy were higher for the AI Explanation version and lower for the Base version for both period dates and ovulation and fertile window (1=very inaccurate to 7=very accurate).

accurate than Base (z ratio=4.61, p<.0001) and marginally more accurate than AI Keywords (z ratio=2.31, p=0.054). They also expected the Base version to be the least accurate for predicting ovulation and fertile window—it was rated less accurate than AI Keywords (z ratio=-3.03, p=0.007) and AI Explanation (z ratio=-4.81, p<0.0001)—and period dates—it was also rated marginally less accurate than AI Keywords (z ratio=-2.26, p=0.061).

These responses suggest that the presence of an AI explanation (e.g., AI Explanation version) positively influences participants' expectations of accuracy. But beyond that, they also indicate that AI descriptions are more positively associated with accuracy even without explanation of how it works (e.g., AI Keywords version). Qualitative results suggest that three main aspects influence this association: (i) general assumptions of AI's capabilities, (ii) previous experiences with fertility apps, and (iii) knowledge about fertility and their bodies.

First, responses suggest that many participants associate AI with scientific results (*"I trust it because KAYA uses AI which is scientific"* - P260) and innovation (*"it uses more advanced code to predict more accurately"* - P70), and those factors are associated with better accuracy (*"it would be the more accurate app, overall. I think that the technology it uses does give it an advantage over an app that does not employ those methods at all"* - P93). These associations convey accuracy to AI descriptions (*"I think things powered by AI are generally pretty accurate"* - P97) and may reinforce individuals' general beliefs and expectations of accuracy in the fertility context.

Second, previous experiences with apps also influence expectations of accuracy. If participants had positive experiences with apps that did not have AI descriptions, participants expected Kaya would be even better because it uses AI: *"I've used apps before that track my period and they have all been rather accurate. I haven't seen them advertise that they use artificial intelligence, so I imagine with this the period tracking would be more accurate with Kaya"* (P88).

In contrast, previous knowledge about fertility and their bodies seem to temper these expectations. Despite the positive association between AI descriptions and accuracy, many participants were aware of its limits. As the next quote illustrates, understanding how factors like stress may affect fertility made some participants doubt the described AI would be very accurate: *"there are some factors that can affect periods, like stress. It might be impossible for Kaya to accurately predict every time"* (P207). However, often participants did not see problems with some level of error: *"I think it couldn't be that far off–matters relating to the body are rarely very precise, and if it's not that far off I am okay with it"* (P4).

5.2 Trust: "How Big the Risk Would Be if Kaya Failed" (P263)

Figure 6 shows the distribution of responses for trust for each of the three main goals. We only observed an influence of descriptions on participants' likelihood of trusting the app for avoiding conception. However, qualitative answers provide more nuance to these results.

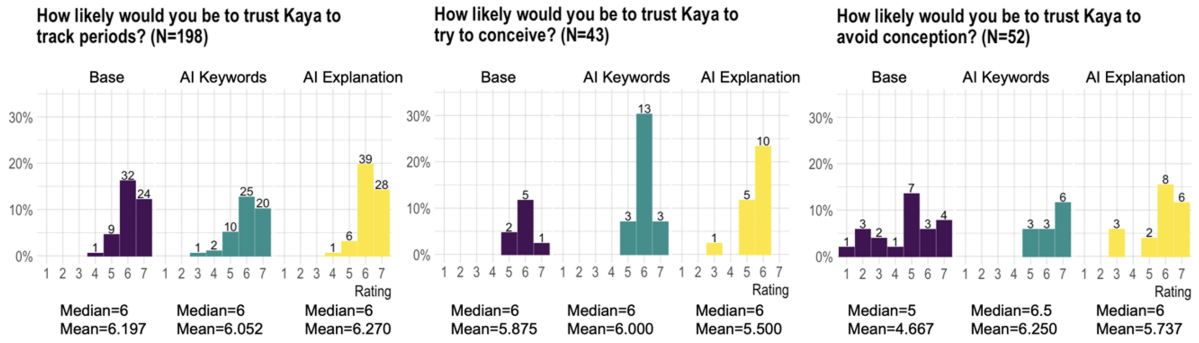

Fig. 6. Participants' trust in the apps: we did not find significant differences on participants' trust in the app for period tracking and ratings were closely aligned. Average answers for AI Keywords and Base versions were higher for trying to conceive but we also did not observe significant differences for this goal. In contrast, we observed a significant difference for participants avoiding conception, with participants considering the AI Keywords version as more trustworthy than the Base version (1=very unlikely to 7=very likely).

We did not observe differences on likelihood of trust for the goal of period tracking (χ^2=1.50, p=0.472). Qualitative responses suggested that descriptions had little influence, since participants often feel that the consequences of inaccuracies are low for this goal: *"There's no harm if there's a mistake in tracking periods, and if you're trying to conceive anyway then the worst thing is you don't get pregnant, so I'm not worried. However, I am not going to trust an app for birth control"* (P4). As the quote illustrates, descriptions and explanations do not appear to matter much for period tracking, but, in contrast, participants considered *"conceiving and birth control are delicate matters"* (P123).

We also did not observe an effect of version on trust among participants trying to conceive (χ^2=3.84, p=0.146). However, trying to conceive qualitative responses were mixed: although some participants understand it may be delicate (*"This is nothing against the app, just my view on how hard it may be or easy it may be to get pregnant depending on the individual"* - P8), others think that errors in the app do not lead to negative consequences in this scenario (*"I would trust it for period tracking and trying to conceive because if the app is wrong, it's no big deal"* - P2). Others even failed to see menstruating when you are actively trying to conceive as a negative consequence that contributes to difficult emotions [16], seeing only unintended pregnancies as negative: *"Period tracking doesn't have a risk factor. If I were trying to conceive then the only consequence would be one I wanted. If I were using Kaya as a birth control method I would need it to be extra reliable"* (P261). Overall, these comments suggest that participants saw trying to conceive as an intermediate goal (higher stakes than period tracking, but lower than avoiding conception) or were unaware of how much tracking may help conception. They also highlight how infertility experiences can be invisible [15].

Finally, different from the other goals, results showed that descriptions had a medium effect (E^2=0.157) on participants' self-reported trust on the app for avoiding pregnancy (χ^2=7.988, p=0.018). The post-hoc test showed that participants with this goal who used the AI Keywords were more likely to trust Kaya than those who used the Base version (Z=2.62, p=0.026), suggesting that the presence of AI terms influenced their answers even without

an explanation. Participants avoiding conception also rated the AI Explanation marginally more trustworthy than the Base version (Z=2.04, p=0.061), suggesting that the presence of an explanation marginally influenced their perceptions when compared to the Base. The mean and median for the AI Keywords was higher (Figure 6 on the right), but we did not observe a significant difference in trust between AI Keywords and AI Explanation (Z=0.82, p=0.413).

However, even with participants rating the AI Keywords version as more trustworthy, they generally reported not believing in using an app as birth control. In the qualitative responses, participants often mentioned they would not trust Kaya (or any app) as their only means to prevent pregnancy (*"It is extremely important to me that I avoid conception, and I would never trust an app for that"* - P86), but would be willing to use it alongside other means: *"I would not use Kaya as a standalone birth control method, but I would use it in addition to other methods"* (P66). These results indicate high tolerance for accuracy if the goal is considered low risk but a low practical trust for known highly consequential tasks even if participants judge the app as accurate: *"I would not base an app on pregnancy because it is not always going to be completely accurate. However, if it gets a period date wrong that is no big deal"* (P19).

In summary, AI descriptions interacted with participants' views of consequences differently for different goals. Because participants saw period tracking and trying to conceive as less consequential, they were more tolerant with predictions accuracy, and we did not observe significant results for trust for these goals. In contrast, avoiding conception results indicate that AI descriptions may be perceived as more trustworthy for highly consequential goals. These results partially align with the accuracy ones. However, we saw an inversion on the preferred versions: while the AI Explanation was rated more positively for accuracy, participants preferred the AI Keywords over Base for trust.

5.3 Willingness to Download and Use: "AI is a Polarizing Topic That Can Scare People Away" (P37)

In general, participants were willing to download all Kaya versions (Figure 7) but preferred the AI Explanation.

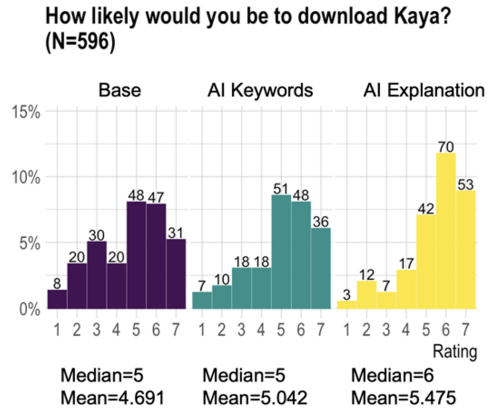

Fig. 7. Participants' willingness to download each app: they were more willing to download the AI Explanation (1=very unlikely to 7=very likely).

Results show a small effect for app version (E^2=0.040) and a medium effect for order (E^2=0.092) on participants' willingness to download Kaya. Participants indicated that they would be more willing to download the AI Keywords (z=2.06, p=0.039, 95%CI 0.02-0.72 higher on a 7-point Likert scale) and the AI Explanation (z=4.72 p<0.001, 95%CI 0.50-1.20 higher on a 7-point Likert scale) than the Base version. The post-hoc test showed

differences between the Base and AI Explanation versions (z ratio= -4.72, p<0.0001) and between the AI Keywords and AI Explanation versions (z ratio=-2.66, p=0.021), indicating that participants are more likely to download the AI Explanation version than the other two. Although the AI Keywords was rated slightly more positively than the Base version (Figure 7), the post-hoc test did not show significant differences (z ratio=-2.06, p=0.097). These results show that unlike explanations, we did not observe a significant effect of the mere presence of AI terms on participants' willingness to download the app.

While results for willingness to download describe general impressions of apps, the analysis for willingness to use (Figure 8) targets each goal separately (i.e., people may be willing to download an app but not to use it for a specific goal). We did not observe the influence of descriptions on willingness to use Kaya for period tracking in the post-hoc test.

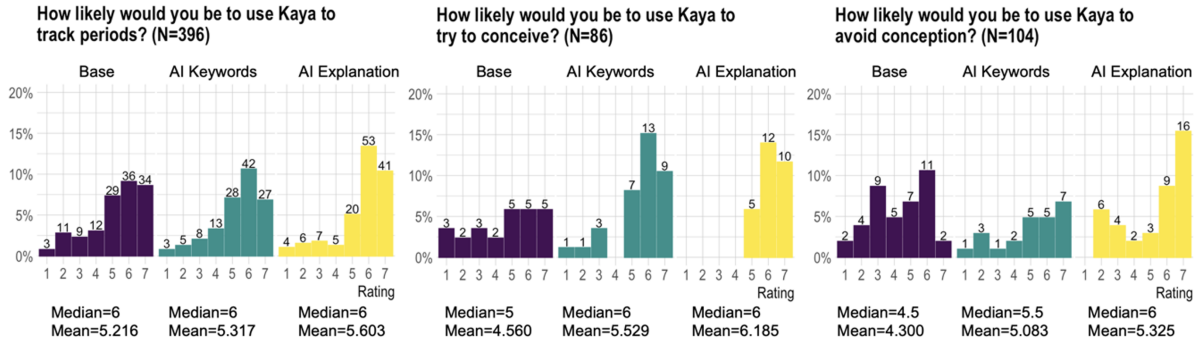

Fig. 8. Participants' willingness to use Kaya for main fertility goals: they were more willing to use the Explanation AI version for all three goals (1=very unlikely to 7=very likely).

For trying to conceive (E^2=0.114) and avoiding conception (E^2=0.0836), the descriptions had a medium influence on participants' willingness to use Kaya (order also had medium effects: E2=0.305 and E2=0.164 respectively), with the AI Explanation being rated higher than the Base version for both goals (trying to conceive: z=2.61, p=0.009, 95%CI 0.37-2.51 higher on a 7-point Likert scale; avoiding conception: z=3.14, p=0.002, 95%CI 0.51-2.16 higher on a 7-point Likert scale). The post-hoc test confirmed that participants were more willing to use the AI Explanation than the Base version for both trying (z-ratio=2.61, p=0.025) and avoiding conception (z-ratio=3.14, p=0.005). AI Keywords' results were not significant for trying to conceive (z=1.27, p=0.204) and the post-hoc test for avoiding conception did not show significant differences. So, for conception goals, participants were more likely to use the AI Explanation than the Base version, suggesting that beyond AI terms, the presence of explanations also influenced participants' perceptions.

Qualitative responses largely align to these results. Nevertheless, they also show a conflict between personalization and mechanization, which characterizes the polarization mentioned by P37: *"AI is a polarizing topic that can scare people away"*. First, when comparing two apps with AI descriptions, participants often preferred the one with explanation (*"I would prefer the second app [AI Explanation] simply because of the more information they share in the description. I feel like these apps are extremely personal and the more information is shared the more likely I will be willing to use it"* - P5). However, when comparing the Base version with the others, particularly the AI Keywords, responses were more split.

The lack of mention of AI may have resulted in the Base reading as *"more user-friendly"* (P16) and *"warmer"* (P44) than AI versions. Many participants believe it would also be *"more personalized and humanistic"* (P31) because it would use only their data (*"it's specifically tailored to my body"* - P58) and it does not use AI (although this is

never stated in the app). These participants often assume AI apps use population data (which is not necessarily true) and would prefer to have predictions based only on their data: *"The first app [Base] relies on data that I input while the second [AI Keywords] relies on an AI to predict and gives a more generalized prediction. The second takes initial data inputted by the user and makes a generalized prediction off of that information and data from other users. (...) [I prefer] The first because it is more based around my personal cycle and takes into account just me as opposed to a general person"* (P73).

Conversely, AI descriptions, particularly with no explanation, sometimes passed a feeling of being *"more impersonal"* (P131), *"pretentious"* (P249), and evoking clinical vocabulary (*"sounds a lot more accurate, it sounds more clinical though"* - P172). To some, AI descriptions imply taking control from users: *"I think with something as intimate and personal as one's period/hormonal changes and stuff, it makes me feel safer to have the perception that I'm in control of this app, not some nameless AI tech"* (P6).

We found that participants did express some concern about how apps might use their data. For example, P5 described they are afraid *"that my information may be stolen"* because of *"news about information being leaked on Period tracking apps and how you can't always trust your data being safe."* For this reason, this participant says they would prefer the Base version over the Keywords version of Kaya. Similarly, P171 says they *"don't feel like it's such a good idea to give out this information to a non-involved third-party so easily."* This was particularly salient in light of conversations around Roe v. Wade (e.g., *"It has made me feel slightly uneasy about tracking my period as I don't want that data to one day be used against me"* – P4). Such concerns were largely orthogonal to the presence of AI. Participants were concerned about data disclosures regardless of the algorithmic approach. However, one participant did express more specific critique to AI models using their data: *"all the talk of machine learning and AI is kind of intimidating and creepy for something I'm putting personal information into, even if it might be slightly more accurate"* (P29).

These negative comments considering AI impersonal, cold, and taking control from users were more present in comparisons between AI Keywords and Base and may help explain why we did not find significant differences between these versions: as P29 comment suggests, the impact of the accuracy attributed to AI descriptions conflicted with negative beliefs and assumptions some participants associated with them. Therefore, it is possible that the ratings for the AI Keywords and Base versions may have ended up similar due to this polarization, with the AI Keywords having benefits around expected accuracy while also feeling impersonal.

In contrast, the AI Explanation was preferred over the Base version in the quantitative and qualitative responses, even though it also advertises using AI: *"I would prefer the second app [AI Explanation over Base] because it is more clear in how it predicts information"* (P77). In general, many participants thought the AI Explanation version was more transparent (*"[I prefer] the first app [AI Explanation] because it is transparent with regards to how they will arrive on the results and what would be required of me"* - P275) and trustworthy (e.g., *"The second app [AI Explanation] provides a description of the prediction system it uses, making the app more trustworthy"* - P181). So, an AI explanation seems to overcome the impersonal and cold feelings AI terms may evoke by implying trustworthiness and a *"scientific background"* (P166). These results show that beyond accuracy, deciding to download and use an app is also influenced by the feelings evoked by the terminologies and descriptions used in the app store page.

6 DISCUSSION

Drawing from research on perceptions of algorithmic recommendations and personal informatics, this study explored how app descriptions influence individuals' perceptions of AI-infused fertility self-tracking apps. While Table 1 summarizes the study's results, in this section, we discuss the nuances of health goals, how literacy and AI assumptions influence individuals' perceptions of AI descriptions, and the caveats of transparency in the context of personal health.

Table 1. Summary of results

Measure	Summary Results
Accuracy	Participants expected the AI Explanation and AI Keywords to be more accurate than the Base version for period dates and fertile window. The AI Explanation was expected to be the most accurate for period dates and the Base version to be the least accurate for ovulation and fertile window. This association between AI descriptions and accuracy was influenced by individuals' assumptions about AI, previous experiences with fertility apps, and knowledge about fertility and their bodies.
Trust	AI descriptions influenced trust for avoiding conception: participants rated the AI Keywords as more trustworthy and the AI Explanation as marginally more trustworthy than the Base version. Despite that, participants reported that they would not use any app as their only birth control method. Participants primarily consider consequences of failure when deciding to trust an app across tracking goals.
Willingness to Download and Use	Participants were more willing to download the AI Explanation version than the others and were more willing to use it than the Base version for trying and avoiding conception, suggesting that the presence of explanation influenced their perceptions. Results for the AI Keywords version and for period tracking were not significant.Responses show a conflict between positive and negative assumptions and beliefs related to AI and how explanations influence this conflict.

6.1 Goals and Their Consequences Are More Important Than Tasks

Previous research has found that individuals perceive algorithmic decisions (including AI-based) as less trustworthy and preferable than human decisions [51–53], especially for tasks considered more delicate or subjective, such as hiring or evaluating workers [11, 51, 58]. In contrast, our findings point that in personal health contexts people desire AI with explanation, associating it with accuracy (especially when descriptions do not provide accuracy measures, as in many commercial apps, which our study emulated). However, our results provide more nuance to these preferences, particularly related to goals.

Our results suggest that goals can be more meaningful or complex than tasks. In the health domain, the same task can be performed for different reasons with different consequences. Fertility is a great example of how this plays out: the task the app executes for both trying and avoiding conception, for example, is to predict fertile days. People even transition through these goals in different life stages, but the task of predicting the fertile window remains fixed. What changes is what users do with predictions and its consequences. Different from previous research [24, 82], expected accuracy was not the main influence on participants' trust intentions; consequences were.

When analyzing trust, we did not observe significant differences for period tracking and trying to conceive, which our participants considered less consequential goals. They even showed some tolerance for errors in predictions for these goals, not considering AI descriptions important. In contrast, for avoiding conception, participants considered both AI versions as more trustworthy than the Base version. This preference on the surface may align with previous research that describes high acceptance and trust on AI systems for any tasks [4, 45]. However, participants explained that they would not trust an app for avoiding conception. Contrasting with quantitative responses, these are more aligned with opposed previous results that show resistance to AI in sensitive tasks [11, 13, 58].

These results suggest a high interest in AI in low-stakes consumer health spaces because people see it as an opportunity for "improving accuracy" of their decisions with minimal negative consequences. Some participants with high-stakes goals were interested in using tools describing AI features, and were particularly accepting

of versions with explanations [66]. However, most expressed some skepticism around blindly trusting them for their highly consequential goals. Instead, they want to use them alongside solutions they know are reliable based on their own experiences [24]. These results highlight the need to design AI for personal health carefully to consider users' different underlying goals and their consequences even within the same health domain and for the same task, since different goals might affect individuals' perceptions of the same description differently. Future studies should consider how this interaction between goals, trust, and accuracy is affected when specific accuracy measures are provided [75] and after users have observed apps' accuracy with use [72].

Beyond that, the fact that participants had different attitudes towards apps depending on their goals even if the task the app did was the same may reflect a standpoint of balanced trust or appropriate reliance [6, 91], which should be encouraged over narrow "acceptance" of AI. However, pregnancy is a clear, important, high-stakes consequence whose outcome or lack thereof can be associated with algorithmic accuracy. Other health contexts may have blurrier impacts when explanations and accuracy information do not directly connect to consequences. These contexts may promote mixed attitudes like the ones we observed for trying to conceive, where people may underestimate the consequences of following app recommendations (e.g., many users did not consider how not conceiving can be heartbreaking to individuals who have this specific goal).

6.2 AI Assumptions and Descriptions Interact and Impact Individuals' Perceptions

Because our study did not provide measures of accuracy and confidence, our results complements previous studies that analyzed these measures [47, 72, 91] by shedding light on assumptions that reveal beliefs and gaps in understanding that feed individuals' initial mental models of AI [91] and may interact with designed descriptions in unexpected ways [40]. Those are important in the health domain because consequences may be critical, as pondered by our participants.

Participants' responses describe their beliefs and assumptions associated with AI, e.g., it uses or does not use their data, it is or is not personalized, it is innovative, it is creepy, etc. These different beliefs and assumptions are strongly connected to participants' AI and health literacy. As Longoni et al. [58] describe, the health domain has intrinsic uncertainty and risks that people are generally not aware of (e.g., many participants were unfamiliar with negative consequences of trying to conceive [16]). Besides, individuals' health literacy varies greatly; low fertility literacy is particularly common [9, 59]. Health and fertility are uncertain domains where people may rely on assumptions to make sense of their experiences [21, 75]. Fertility self-tracking sees the body as its object and subject at the same time, a complex process that *"involves negotiating and making sense of external sources of information in conjunction with bodily experience"* [41]. With the growing commercialization of AI, another layer of uncertainty is added to this domain since AI demands specific literacy [2, 26, 47, 72]. In general, our results suggest that how participants made sense of apps' descriptions was influenced by assumptions related to dominant AI discourses [36], a widespread lack of knowledge about fertility, and their own health and technology literacy levels.

This combination of assumptions and literacy led to two polarized views of AI: one based on positive perspectives of innovation and accuracy and another one based on negative perspectives of loss of control and lack of personalization. This polarization interacted with app descriptions in different ways. For example, our quantitative results for trust may indicate that, without comparing the AI Keywords with others, many participants assume its accuracy would be improved (positive perspective), although the description does not provide accuracy measures and AI may not provide more accurate results [43, 56]. However, after seeing the Base version, qualitative responses suggest that the negative beliefs and assumptions about AI carried more weight and many preferred to use an app that felt warmer and more personalized, despite the general expectation of accuracy (negative perspective). This polarization likely influenced our findings regarding willingness to download and use, which did not surface significant differences between these versions.

In comparison, when AI descriptions included an explanation, the negative perspective waned, and people preferred the app that they perceived as more transparent even if the description was colder and impersonal. Participants who used the AI Explanation may have had second thoughts at first because the more detailed explanation may have drawn their attention to uncertainty [6]. Nevertheless, when they compared it with the other versions, responses suggest that their opinions became more positive because the AI Explanation was largely considered more transparent and preferred over the others.

These results illustrate how individuals' AI assumptions may affect how users understand apps' descriptions. First, consumers may choose to use and trust an app that has warmer descriptions that imply personalization but does not explain how the app works, than one that provides explanations using more technical terms. People's trust and use over time will be influenced by apps' perceived performance and accuracy when users in fact use the system [72]. However, the behavior of choosing an app that implies personalization might give apps an excuse to collect more data under the guise that it will be used for predictions, which may potentially have dire consequences for users' privacy in such a sensitive context [67, 77, 83]. Second, aligned with previous studies [6, 66], our results show that participants want more meaningful information and transparency. In our case, using only vague technical terms (AI Keywords) may have sounded performative, only a marketing strategy to imply technological innovation and attract users [81]. Providing more information may give users a sense of being part of the process [66], instead of patronizing them with void technology words. However, previous research found that explanations increased reliance on AI recommendations regardless of its correctness [6]. So, instead of inspiring appropriate trust, algorithmic explanations could lead users to uncritically rely on AI recommendations that may be detrimental to their health, especially when accuracy (stated and perceived) does not directly map to consequences. These trade-offs call attention to the possibility of using deceptive design solutions in explanations as dark patterns, which are ultimately not designed to benefit users [14]. They also call for more attention on designing AI for personal health, particularly broadening our view on transparency.

6.3 AI in Personal Health: The Limitations of Transparency

Previous studies suggest a need to aid health consumers in understanding both AI [82] and PI [7] metrics and features so they can better evaluate apps' recommendations. Additionally, our study showed that participants preferred AI descriptions with explanation. However, transparency alone may not be enough and prior research has called attention to its limitations [3, 36, 91]. Of particular interest in this study: transparency does not necessarily build trust and it can privilege seeing over understanding [3].

As Ananny and Crawford explain, transparency does not necessarily improve trust because people decide to trust systems depending on different factors [3]. In this study, many factors influenced trust beyond accuracy; perceived consequence of a failure was the biggest one. Personalization, perceived control, and friendliness were also considered by participants in different ways and factors outside of the algorithm's workings can also be influential. For instance, the recent US Supreme Court ruling overturning abortion rights generated discussions about the use of fertility tracking apps [67, 83] and influenced many participants' perceptions and trust on these tools, regardless of AI use. So, US individuals may distrust these apps more not because of how they generate predictions but because they are afraid their data can be used against them.

Transparency can also privilege seeing over understanding because learning about a complex system requires more than seeing inside its "black-box" [3]: it is necessary to interact with the system to understand how it works in relation to the environmental and social contexts where it is embedded, so one can challenge the system when necessary. Seeing how recommendations are generated does not necessarily mean understanding the health processes and the potential consequences of health decisions. If people see AI explanations but do not have enough health and AI literacy, they may not know the potential consequences of relying on them, particularly in

early use. In this context, users can be also prone to accept placebo explanations or even to be manipulated by dark patters used in explanation design [14].

Therefore, to design AI systems for personal health that inspire appropriate trust and reliance, we need to focus on aspects of the social context where the tool is situated alongside the explanation. First, it is important to consider participants' health literacy [24]: novice users may be more influenced by AI descriptions, while experienced users may have already developed their understanding of how their bodies [87] and technologies work. This difference is especially important for fertility because of its complexity and stigma [1, 17], but individuals who endeavor to learn about it often become experts on their bodies over time [18]. For the latter, developing and maintaining appropriate trust on technologies may be more natural because of their developed expertise. But for novices or people who lack fertility expertise in general, AI explanations may be more crucial [91] and may even define what fertility is to them [18, 41]. Future research should investigate how bodily and fertility expertise (perceived and tested) influence individuals' attitudes and how explanations can support novices to build literacy so they can develop appropriate trust. Similarly, it is important to analyze how expertise and explanations intertwine with privacy concerns and risks and their relationships with dark patterns of design [14], especially considering the differential vulnerabilities that individuals may have [62].

Second, the influence of AI assumptions needs to be further analyzed to understand its consequences. Similar to Grill and Andalibi's [36] work on emotion recognition, identifying and recognizing individuals' assumptions and folk theories about AI in consumer health technologies sheds light on what users consider important and able to support them in developing knowledge about such systems. It also supports an analysis of emotional effects of these technologies on individuals' experiences and the social and political consequences of these assumptions [41, 42], which can generate insights on how to tackle them through explanations (if possible) and how to combat dark patterns in explanation design [14]. Furthermore, findings from Explainable AI [8, 90] research need to be further translated to PI to produce explanations that are understandable, useful, and consider non-experts assumptions and literacy to support them in practical, real-life situations [13, 82] that have more emotional and critical consequences than simply influencing technology acceptance. For example, our results suggest that including the potential consequences of decisions for different goals is as important as accuracy and explanations about algorithmic processes. Besides developing meaningful and situated explanations, we also need to investigate further how people understand them and the algorithmic recommendations they explain [66] and how these understandings influence their embodied lived experience [41] and their privacy [62, 77]. For example, future research should extend this study including a white-box version [13], analyzing longer use or removal [41], and investigating the influence of AI descriptions on individuals' understanding of fertility and AI algorithms and its relation to data privacy.

7 LIMITATIONS

Like similar studies [4, 45, 51, 53, 58], our methodology presents some limitations. First, we acknowledge that there are differences between measuring AI perceptions based on descriptions and based on apps' actual performance [64]. Although Kaya allowed participants to briefly use the app, lived experience is necessary for people to develop better understanding of the workings and limitations of algorithms [50]. However, previous research has shown that scenario-based studies can help understand users' perceptions [4, 45, 51, 53, 58], which impact how accepting they are of technology [19, 82]. Therefore, this methodology proved to be suitable to investigate our research question.

Second, we developed Kaya versions to investigate how AI terms and the presence of AI explanations influence participants' perceptions of fertility apps. Our intention was not to investigate if people prefer non-AI algorithms or AI algorithms. So none of our versions explicitly state they followed a non-AI approach, as many apps also do not do [31]. Therefore, our results may not extend to preferences towards non-AI algorithms, since an app

version without AI but providing a more detailed explanation of its calculations could wield different perceptions in participants – an approach that is a valuable opportunity for future work. Besides, similar to Newn et al. [66], we did not intend to evaluate different styles of explanation, but the presence of them [66]. So, we used only one explanation style through our AI Explanation version. We also found that the description for the Base version might have reinforced the belief that users' control over their data would be higher. Yet, the three versions used in this study provide valuable insights into how individuals' perceptions interact with their assumptions when they evaluate and choose fertility tracking apps.

Third, while asking participants to compare the version they walked through to a second app deepened our ability to understand how descriptions influenced perspectives, it also introduced a strong order effect. In general, participants rated the second app more negatively than the first. We have a few possible explanations for that. First, a priming effect: after seeing the second app, participants may have reflected more and become more critical about the second app. Second, participants used the first app while the second they only saw the app store page. Therefore, they knew more about the first app and there was more room for confusion for the second one. And third, a fatigue effect since the comparison happens towards the end of the survey. However, we were still able to answer our research question effectively because order was randomized and therefore controlled for. Finally, unlike the app version, we did not control for participants' primary goals in recruitment, so our results replicate general population interests [30], which may have impacted the goal analysis because sample sizes for trying and avoiding conception were small.

Finally, other limitations are related to the participants' pool. We recruited from Prolific and thus our participants reflect the population who use this platform and differed from the US population in a few key aspects, particularly the ratio of white and Black participants, household income, and education attainment. These differences may correlate with higher mistrust of medical professionals [53], technology acceptance, and fertility knowledge. In addition, fewer participants had children or described having irregular cycles, factors that may influence their beliefs of how accurately AI could predict their cycles or their trust in these tools in general. Future studies should investigate how participants' demographics might influence their perceptions. We also highlight that other stakeholders such as male partners are often important actors in fertility journeys [18, 25], may use these apps for various reasons, and that a few fertility apps even include features to support shared experiences [38]. However, our results do not extend to these users and studying the perceptions of non-pregnant partners is another important direction for future work. Finally, US social, cultural, economic, and political aspects might have influenced the results [45]. For example, almost half participants described that the overturning of abortion rights [67, 83] affected their thoughts on fertility self-tracking apps, if only slightly: many of them are afraid that people's self-tracked data could be sold or used in lawsuits, while several others described believing that these apps are more important than ever because people will need more control over their bodies with the scenario of reduced reproductive rights. These factors may have directly influenced participants' perceptions. Overall, we think it led to slightly greater skepticism about these sorts of apps in general, but not specifically about perceptions of AI features. While we opted to describe Kaya as only storing data locally, it has been reported that this is not the case of most fertility apps [62, 77]. Given potential differences in how AI v. non-AI apps might handle private data (cloud- v. local, for example) and potential risks associated with intimate health-related data, the relationship between data use, privacy concerns, and social contexts should be more deeply considered in future studies.

8 CONCLUSION

In examining the influence of descriptions on individuals' perceptions of fertility self-tracking, we found that participants associate AI descriptions with increased accuracy and are more willing to trust, download, and use apps that advertise AI. Despite this general preference for tools describing AI, participants reported that they

would not blindly trust them for avoiding conception: beyond accuracy, the consequences of a failure were more important to define their trust. Our results also show a polarization between positive and negative assumptions and beliefs related to AI, with some individuals praising its accuracy while others thinking AI is impersonal and cold. Based on these results, we discussed the nuances of health goals and their relation to the perceptions of app descriptions, how literacy and AI assumptions influence individuals' perceptions of AI descriptions, and the limitations of transparency in the context of AI for personal health.

ACKNOWLEDGMENTS

We thank our participants for their sincere participation. We thank Thu Anh Huynh for developing Kaya's design and Marawin Chheang, Edward Wu, Neal Khodaskar, Donggun Lee, Joonyoung Park, Ye Lin Jeong for participating in different stages of Kaya's prototype development. We also thank Mustafa Hussain for his support and feedback on earlier stages of this project. This research was supported in part by the 2020 Microsoft Dissertation Grant from Microsoft Research, the 2020 Exploration Award from the University of California, Irvine, Donald Bren School of Information and Computer Science, and the National Science Foundation under award IIS-2237389.

REFERENCES

[1] Teresa Almeida, Rob Comber, and Madeline Balaam. 2016. HCI and Intimate Care as an Agenda for Change in Women's Health. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 2599–2611.

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

[3] Mike Ananny and Kate Crawford. 2018. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *new media & society* 20, 3 (2018), 973–989.

[4] Theo Araujo, Natali Helberger, Sanne Kruikemeier, and Claes H De Vreese. 2020. In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & society* 35 (2020), 611–623.

[5] Boris Babic, Sara Gerke, Theodoros Evgeniou, and I Glenn Cohen. 2021. Direct-to-consumer medical machine learning and artificial intelligence applications. *Nature Machine Intelligence* 3, 4 (2021), 283–287.

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

[7] Marit Bentvelzen, Jasmin Niess, and Paweł W Woźniak. 2023. Designing Reflective Derived Metrics for Fitness Trackers. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 4 (2023), 1–19.

[8] Or Biran and Courtenay Cotton. 2017. Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, Vol. 8. 8–13.

[9] Laura Bunting, Ivan Tsibulsky, and Jacky Boivin. 2013. Fertility knowledge and beliefs about fertility treatment: findings from the International Fertility Decision-making Study. *Human reproduction* 28, 2 (2013), 385–397.

[10] US Census Bureau. 2020. 2020 census illuminates racial and ethnic composition of the country.

[11] Noah Castelo, Maarten W Bos, and Donald R Lehmann. 2019. Task-dependent algorithm aversion. *Journal of Marketing Research* 56, 5 (2019), 809–825.

[12] CDC. 2022. Women's Reproductive Health. https://www.cdc.gov/reproductivehealth/womensrh/index.htm

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

[14] Michael Chromik, Malin Eiband, Sarah Theres Völkel, and Daniel Buschek. 2019. Dark Patterns of Explainability, Transparency, and User Control for Intelligent Systems.. In *IUI workshops*, Vol. 2327.

[15] Mayara Costa Figueiredo. 2021. *Data Work and Data Tracking Technologies in Fertility Care: A Holistic Approach.* University of California, Irvine.

[16] Mayara Costa Figueiredo, Clara Caldeira, Elizabeth Victoria Eikey, Melissa Mazmanian, and Yunan Chen. 2018. Engaging with health data: The interplay between self-tracking activities and emotions in fertility struggles. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–20.

[17] Mayara Costa Figueiredo, Clara Caldeira, Tera L Reynolds, Sean Victory, Kai Zheng, and Yunan Chen. 2017. Self-tracking for fertility care: collaborative support for a highly personalized problem. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–21.

[18] Mayara Costa Figueiredo and Yunan Chen. 2021. Health data in fertility care: an ecological perspective. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–17.

[19] Mayara Costa Figueiredo, Thu Huynh, Anna Takei, Daniel A Epstein, and Yunan Chen. 2021. Goals, life events, and transitions: examining fertility apps for holistic health tracking. *JAMIA open* 4, 1 (2021), ooab013.

[20] Mayara Costa Figueiredo, H Irene Su, and Yunan Chen. 2020. AN EVALUATION OF COMMERCIAL FERTILITY APPS: ALGORITHMIC PREDICTIONS AND USERS'PERCEPTIONS. *Fertility and Sterility* 114, 3 (2020), e552–e553.

[21] Mayara Costa Figueiredo, H Irene Su, and Yunan Chen. 2021. Using data to approach the unknown: Patients' and healthcare providers' Data practices in fertility challenges. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–35.

[22] John W Creswell and Vicki L Plano Clark. 2017. *Designing and conducting mixed methods research*. Sage publications.

[23] Abdallah S Daar and Zara Merali. 2002. Infertility and social suffering: the case of ART in developing countries. *Current practices and controversies in assisted reproduction* 15 (2002), 21.

[24] Pooja M Desai, Elliot G Mitchell, Maria L Hwang, Matthew E Levine, David J Albers, and Lena Mamykina. 2019. Personal health oracle: Explorations of personalized predictions in diabetes self-management. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[25] Patel Dilisha, Blandford Ann, Warner Mark, Shawe Jill, and Stephenson Judith. 2019. I Feel like Only Half a Man": Online Forums as a Resource for Finding a" New Normal" for Men Experiencing Fertility Issues. 3. *Proc. ACMHum.-Comput. Interact* 3 (2019).

[26] Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. 2017. UX design innovation: Challenges for working with machine learning as a design material. In *Proceedings of the 2017 chi conference on human factors in computing systems*. 278–288.

[27] Jane Dreaper. 2016. Women warned about booming market in period tracker apps. https://www.bbc.com/news/health-37013217#

[28] Marguerite Duane, Alison Contreras, Elizabeth T Jensen, and Amina White. 2016. The performance of fertility awareness-based method apps marketed to avoid pregnancy. *The Journal of the American Board of Family Medicine* 29, 4 (2016), 508–511.

[29] Daniel A Epstein, Clara Caldeira, Mayara Costa Figueiredo, Xi Lu, Lucas M Silva, Lucretia Williams, Jong Ho Lee, Qingyang Li, Simran Ahuja, Qiuer Chen, et al. 2020. Mapping and taking stock of the personal informatics literature. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–38.

[30] Daniel A Epstein, Nicole B Lee, Jennifer H Kang, Elena Agapie, Jessica Schroeder, Laura R Pina, James Fogarty, Julie A Kientz, and Sean Munson. 2017. Examining menstrual tracking to inform the design of personal informatics tools. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 6876–6888.

[31] Mozilla Foundation. [n. d.]. *Privacy Not Included: A Buyer's Guide for Connected Products. https://foundation.mozilla.org/en/privacynotincluded/?search=period

[32] Alexander Freis, Tanja Freundl-Schütt, Lisa-Maria Wallwiener, Sigfried Baur, Thomas Strowitzki, Günter Freundl, and Petra Frank-Herrmann. 2018. Plausibility of menstrual cycle apps claiming to support conception. *Frontiers in Public Health* 6 (2018), 98.

[33] A Frost and R Sullivan. 2021. Femtech–Time for a digital revolution in the women's health market.

[34] Elliot G. Mitchell, Elizabeth M. Heitkemper, Marissa Burgermaster, Matthew E. Levine, Yishen Miao, Maria L. Hwang, Pooja M. Desai, Andrea Cassells, Jonathan N. Tobin, Esteban G. Tabak, et al. 2021. From reflection to action: combining machine learning with expert knowledge for nutrition goal recommendations. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–17.

[35] Sara Gerke, Timo Minssen, and Glenn Cohen. 2020. Ethical and legal challenges of artificial intelligence-driven healthcare. In *Artificial intelligence in healthcare*. Elsevier, 295–336.

[36] Gabriel Grill and Nazanin Andalibi. 2022. Attitudes and folk theories of data subjects on transparency and accuracy in emotion recognition. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–35.

[37] Kevin Hamilton, Karrie Karahalios, Christian Sandvig, and Motahhare Eslami. 2014. A path to understanding the effects of algorithm awareness. In *CHI'14 extended abstracts on human factors in computing systems*. 631–642.

[38] Josie Hamper. 2022. A fertility app for two? Women's perspectives on sharing conceptive fertility work with male partners. *Culture, health & sexuality* 24, 12 (2022), 1713–1728.

[39] Victoria Hollis, Alon Pekurovsky, Eunika Wu, and Steve Whittaker. 2018. On being told how we feel: how algorithmic sensor feedback influences emotion perception. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–31.

[40] Sarah Homewood, Laurens Boer, and Anna Vallgårda. 2020. Designers in white coats: deploying ovum, a fertility tracking device. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[41] Sarah Homewood, Amanda Karlsson, and Anna Vallgårda. 2020. Removal as a method: A fourth wave HCI approach to understanding the experience of self-tracking. In *Proceedings of the 2020 ACM designing interactive systems conference*. 1779–1791.

[42] Sarah Homewood and Anna Vallgårda. 2020. Putting phenomenological theories to work in the design of self-tracking technologies. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. 1833–1846.

[43] Sarah Johnson, Lorrae Marriott, and Michael Zinaman. 2018. Can apps and calendar methods predict ovulation with accuracy? *Current Medical Research and Opinion* 34, 9 (2018), 1587–1594.
[44] Ingrid Johnston-Robledo and Joan C Chrisler. 2020. The menstrual mark: Menstruation as social stigma. *The Palgrave handbook of critical menstruation studies* (2020), 181–199.
[45] Shivani Kapania, Oliver Siy, Gabe Clapper, Azhagu Meena SP, and Nithya Sambasivan. 2022. " Because AI is 100% right and safe": User attitudes and sources of AI authority in India. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–18.
[46] M Keierleber. 2022. Young and depressed? Try Woebot! The rise of mental health chatbots in the US. *The Guardian* (2022).
[47] Rafal Kocielnik, Saleema Amershi, and Paul N Bennett. 2019. Will you accept an imperfect ai? exploring designs for adjusting end-user expectations of ai systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.
[48] J Koetsier. 2020. AI-driven fitness: Making gyms obsolete. *Forbes* (2020).
[49] Markus Langer, Tim Hunsicker, Tina Feldkamp, Cornelius J König, and Nina Grgić-Hlača. 2022. "Look! It'sa computer program! It's an algorithm! It's AI!": Does terminology affect human perceptions and evaluations of algorithmic decision-making systems?. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–28.
[50] Jong Ho Lee, Jessica Schroeder, and Daniel A Epstein. 2021. Understanding and supporting self-tracking app selection. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (2021), 1–25.
[51] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1 (2018), 2053951718756684.
[52] Min Kyung Lee and Su Baykal. 2017. Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division. In *Proceedings of the 2017 acm conference on computer supported cooperative work and social computing*. 1035–1048.
[53] Min Kyung Lee and Katherine Rich. 2021. Who is included in human perceptions of AI?: Trust and perceived fairness around healthcare AI and cultural mistrust. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–14.
[54] Ian Li, Anind Dey, and Jodi Forlizzi. 2010. A stage-based model of personal informatics systems. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 557–566.
[55] Kathy Li, Iñigo Urteaga, Amanda Shea, Virginia J Vitzthum, Chris H Wiggins, and Noémie Elhadad. 2022. A predictive model for next cycle start date that accounts for adherence in menstrual self-tracking. *Journal of the American Medical Informatics Association* 29, 1 (2022), 3–11.
[56] Bo Liu, Shuyang Shi, Yongshang Wu, Daniel Thomas, Laura Symul, Emma Pierson, and Jure Leskovec. 2019. Predicting pregnancy using large-scale data from a women's health tracking mobile application. In *The World Wide Web Conference*. 2999–3005.
[57] Jennifer M Logg, Julia A Minson, and Don A Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (2019), 90–103.
[58] Chiara Longoni, Andrea Bonezzi, and Carey K Morewedge. 2019. Resistance to medical artificial intelligence. *Journal of Consumer Research* 46, 4 (2019), 629–650.
[59] Lisbet S Lundsberg, Lubna Pal, Aileen M Gariepy, Xiao Xu, Micheline C Chu, and Jessica L Illuzzi. 2014. Knowledge, attitudes, and practices regarding conception and fertility: a population-based survey among reproductive-age United States women. *Fertility and sterility* 101, 3 (2014), 767–774.
[60] Salvatore S Mangiafico. [n. d.]. Kruskal–wallis test. https://rcompanion.org/handbook/F_08.html
[61] Kevin McElrath and Michael Martin. 2021. Bachelor's Degree Attainment in the United States: 2005 to 2019. American Community Survey Briefs. ACSBR-009. *US Census Bureau* (2021).
[62] Maryam Mehrnezhad and Teresa Almeida. 2021. Caring for intimate data in fertility technologies. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–11.
[63] S Merchant. 2021. The best chatbots for behavioral health. AIM.
[64] Christian Meurisch, Cristina A Mihale-Wilson, Adrian Hawlitschek, Florian Giger, Florian Müller, Oliver Hinz, and Max Mühlhäuser. 2020. Exploring user expectations of proactive AI systems. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–22.
[65] Michelle L Moglia, Henry V Nguyen, Kathy Chyjek, Katherine T Chen, and Paula M Castaño. 2016. Evaluation of smartphone menstrual cycle tracking applications using an adapted APPLICATIONS scoring system. *Obstetrics & Gynecology* 127, 6 (2016), 1153–1160.
[66] Joshua Newn, Ryan M Kelly, Simon D'Alfonso, and Reeva Lederman. 2022. Examining and Promoting Explainable Recommendations for Personal Sensing Technology Acceptance. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 1–27.
[67] Sarah Perez. 2022. Consumers swap period tracking apps after Roe v. Wade ruling. https://techcrunch.com/2022/06/27/consumers-swap-period-tracking-apps-in-search-of-increased-privacy-following-roe-v-wade-ruling/
[68] W Nicholson Price and I Glenn Cohen. 2019. Privacy in the age of medical big data. *Nature medicine* 25, 1 (2019), 37–43.
[69] Prolific. [n. d.]. Prolific · quickly find research participants you can trust. https://www.prolific.com/

[70] Mashfiqui Rabbi, Min Hane Aung, Mi Zhang, and Tanzeem Choudhury. 2015. MyBehavior: automatic personalized health feedback from user behaviors and preferences using smartphones. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. 707–718.
[71] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as mechanisms for supporting algorithmic transparency. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–13.
[72] Amy Rechkemmer and Ming Yin. 2022. When confidence meets accuracy: Exploring the effects of multiple performance indicators on trust in machine learning models. In *Proceedings of the 2022 chi conference on human factors in computing systems*. 1–14.
[73] Johnny Saldaña. 2021. *The coding manual for qualitative researchers*. sage.
[74] Morgan Klaus Scheuerman, Katta Spiel, Oliver L Haimson, Foad Hamidi, and Stacy M Branham. 2020. HCI guidelines for gender equity and inclusivity. *UMBC Faculty Collection* (2020).
[75] Hanna Schneider, Julia Wayrauther, Mariam Hassib, and Andreas Butz. 2019. Communicating uncertainty in fertility prognosis. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–11.
[76] Elizabeth Schulze. 2019. 40% of A.I. start-ups in Europe have almost nothing to do with A.I., research finds. https://www.cnbc.com/2019/03/06/40-percent-of-ai-start-ups-in-europe-not-related-to-ai-mmc-report.html
[77] Laura Shipp and Jorge Blasco. 2020. How private is your period?: A systematic analysis of menstrual app privacy policies. *Proc. Priv. Enhancing Technol.* 2020, 4 (2020), 491–510.
[78] Emily A Shrider, Melissa Kollar, Frances Chen, Jessica Semega, et al. 2021. Income and poverty in the United States: 2020. *US Census Bureau, Current Population Reports* P60-273 (2021).
[79] Leon Speroff and Marc A Fritz. 2005. *Clinical gynecologic endocrinology and infertility*. lippincott Williams & wilkins.
[80] Aaron Springer, Victoria Hollis, and Steve Whittaker. 2017. Dice in the black box: User experiences with an inscrutable algorithm. In *2017 AAAI Spring Symposium Series*.
[81] Mary Summer Starling, Zosha Kandel, Liya Haile, and Rebecca G Simmons. 2018. User profile and preferences in fertility apps for preventing pregnancy: an exploratory pilot study. *Mhealth* 4 (2018).
[82] Zhaoyuan Su, Mayara Costa Figueiredo, Jueun Jo, Kai Zheng, and Yunan Chen. 2020. Analyzing description, user understanding and expectations of ai in mobile health applications. In *AMIA Annual Symposium Proceedings*, Vol. 2020. American Medical Informatics Association, 1170.
[83] Rina Torchinsky. 2022. How period tracking apps and data privacy fit into a post-Roe v. Wade climate. *National Public Radio. URL https://www. npr. org/2022/05/10/1097482967/roev-wade-supreme-court-abortion-period-apps* (2022).
[84] Inigo Urteaga, Kathy Li, Amanda Shea, Virginia J Vitzthum, Chris H Wiggins, and Noémie Elhadad. 2021. A generative modeling approach to calibrated predictions: a use case on menstrual cycle length prediction. In *Machine Learning for Healthcare Conference*. PMLR, 535–566.
[85] James Vincent. 2019. Forty percent of "ai startups" in Europe don't actually use AI, claims report. https://www.theverge.com/2019/3/5/18251326/ai-startups-europe-fake-40-percent-mmc-report
[86] Jeffrey Warshaw, Tara Matthews, Steve Whittaker, Chris Kau, Mateo Bengualid, and Barton A Smith. 2015. Can an Algorithm Know the" Real You"? Understanding People's Reactions to Hyper-personal Analytics Systems. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 797–806.
[87] Samantha A Whitman, Kathleen H Pine, Bjorg Thorsteinsdottir, Paige Organick-Lee, Anjali Thota, Nataly R Espinoza Suarez, Erik W Johnston, and Kasey R Boehmer. 2021. Bodily experiences of illness and treatment as information work: The case of chronic kidney disease. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–28.
[88] Rayoung Yang, Eunice Shin, Mark W Newman, and Mark S Ackerman. 2015. When fitness trackers don't'fit' end-user difficulties in the assessment of personal tracking device accuracy. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 623–634.
[89] Yue You, Yubo Kou, Xianghua Ding, and Xinning Gui. 2021. The medical authority of AI: A study of AI-enabled consumer-facing health technology. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.
[90] Yongfeng Zhang, Xu Chen, et al. 2020. Explainable recommendation: A survey and new perspectives. *Foundations and Trends® in Information Retrieval* 14, 1 (2020), 1–101.
[91] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 295–305.