Powered by AI: Analyzing the Influence of Artificial Intelligence Claims in Fertility Self-Tracking

Mayara Costa Figueiredo
University of California, Irvine
Irvine, CA 92697, USA
mcostafi@uci.edu

Mustafa I. Hussain
University of California, Irvine
Irvine, CA 92697, USA
mihussai@uci.edu

Elizabeth A. Ankrah
University of California, Irvine
Irvine, CA 92697, USA
eankrah@uci.edu

Yunan Chen
University of California, Irvine
Irvine, CA 92697, USA
yunanc@ics.uci.edu

CHI 2020 Extended Abstracts, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6819-3/20/04.

Abstract

In the past few years, there has been a proliferation of personal health applications, "powered by" artificial intelligence (AI) and positioned to help individuals make informed health decisions based on their personal data. In particular, fertility self-tracking is an area in which the use of direct-to-consumer AI is rising, and people are increasingly using these tools to seek or avoid conception. While these applications may have the potential to engage and empower laypersons to better understand their fertility, they often act as "black boxes," offering little transparency as to how their algorithms work, and what personal data are used to make predictions. In this paper, we outline an ongoing study on how descriptions of "data-driven AI" influence users' understanding of algorithm feedback about their personal data. Methods include an evaluation of commercial fertility self-tracking tools and an experiment to examine how users' interpretation and trust change based on algorithms descriptions.

Author Keywords

Fertility self-tracking; direct-to-consumer AI; algorithm black box.

CSS Concepts

• **Human-centered computing~Human computer interaction (HCI) • Social and professional topics~Women** • Applied computing~Consumer health

Introduction

Artificial Intelligence (AI) algorithms are increasingly common in healthcare. Most often, AI has been proposed or used in healthcare to support physicians and hospital administration, for example, to identify sepsis early [15], or to predict patients' risk of hospital readmission [19] based on psychosocial factors. However, in the past few years, there has been a proliferation of direct-to-consumer AI: personal health applications which use artificial intelligence to help individuals make informed health decisions based on their personal data. These applications cover a wide spectrum of lifestyle choices and health conditions, from diet control and physical exercises to mental illness. While these systems may have potential to engage and empower laypersons to take better control of their health, such applications often appear as "black boxes" to end users, offering very 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.

Interpretation and Trust in Algorithmic Feedback

HCI researchers have previously described the challenges users face as they attempt to understand and interpret the inner workings and outward behavior of complex computer systems [1,6,12]. Recently, researchers have examined how people interpret and trust algorithmic output [1,6,12,18,20]. Yang et al. [20] have noted that "the black-box nature" of algorithms can "inhibit users from understanding" how they work. This can be problematic when users "over-trust" automated systems [5]. For example, Hollis et al. [12] investigated how algorithmic feedback may influence users' evaluations of their own emotions. They reported that users may defer to an algorithm's classification of their own emotional experience over their own personal judgment of that experience. These results suggest that users may place an inappropriately high level of trust in "black-boxed" algorithms [17]. Another study explored different strategies to explain algorithmic feedback to non-expert end-users [6]. The authors suggested that very detailed explanations on internal algorithm procedures (i.e., a "white-box" approach) could contribute to information overload, also posing barriers for non-expert end-users to understand how their technologies or devices work.

With the popularity of consumer-facing AI, it is crucial for us to understand how AI might influence non-expert users' trust and usages of such systems, especially when making decisions based on these tools' feedback can bring critical consequences to users' lives—for example, in self-tracking for fertility.

AI in Fertility Self-Tracking

Fertility self-tracking is an area that has seen an increased use of AI, particularly in mobile apps marketed directly to consumers. These *fertility apps* allow people to collect diverse health indicators potentially related to their fertility cycles—such as period dates and other physical and emotional data—and provide feedback for users, including predictions for periods, ovulation, and fertility window. Fertility tracking is becoming a popular market, with more than

Preliminary Experiment Design

We will use a within-subject experiment: the same participant will see both versions of the simulated app. The order of the versions will be randomized to reduce bias.

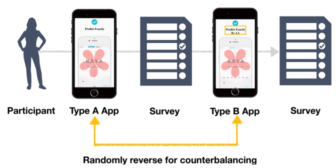

Figure 1: Preliminary design.

Participants will be guided towards a simulation of a fertility app, starting from downloading the app from the app store, inputting data and analyzing its main visualizations. Figure 2

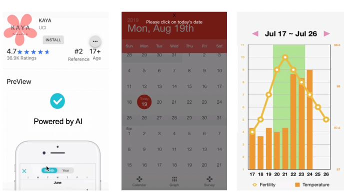

Figure 2: Drafts of prototype's screens—app store page, calendar, and graph visualizations.

300 fertility apps currently available. In 2016, these apps were downloaded around 200 million times worldwide [9].

Many people are using these tools, and trusting their predictions, to either achieve or avoid a pregnancy—goals that can be very emotionally-loaded [7] and potentially life-changing. Besides, many people have only low to intermediate knowledge about fertility [3,4,11,14]. These aspects make the use of AI in fertility tracking a critical case to be studied.

Currently, many fertility apps are claimed to use 'artificial intelligence,' 'smart algorithms,' or 'machine learning.' These terms are often used interchangeably, and they appear to be intended to convey a high level of predictive accuracy. However, most apps do not explain their fertility algorithms or how they generate predictions [8,10,13,16]. It is also not clear what data, among the diverse health indicators collected via these apps, are used to generate fertility predictions. Further, vague descriptions of artificial intelligence algorithms may imply objectivity, credibility, or access to privileged information, influencing peoples' beliefs concerning their capabilities [12,17].

In this study we want to analyze how descriptions of AI might influence users' interpretation and trust in fertility self-tracking apps.

Methods

This paper outlines an ongoing study focusing on how AI claims influence users' understanding of algorithm feedback about their personal data in the context of fertility self-tracking. We intend to address the following research questions:

1. How do users interpret outputs of fertility algorithms, such as the visualizations?
2. How AI descriptions influence users' decisions on adopting the system and trusting their results?

To approach these questions, we will first perform an evaluation of commercial fertility self-tracking tools using published guidelines for human-AI interaction [2]. Second, we designed an experiment to examine how users' trust and interpretation of apps' feedback differ based on AI claims. We prototyped an app walkthrough experience for two versions of a fertility app. Version A promises "artificial intelligence," "machine learning," and "data-driven results." Version B only provides the predictions. The rest of the prototype looks exactly the same.

Participants will be randomly assigned to use the different versions of the prototype. For each version, they will be led through the process of reading the App Store page of the app, downloading the app, inputting fertility-related data, and analyzing a calendar and a graph visualization.

After participants complete each app version, they will be asked to answer a survey covering the following: (1) Their interpretation of visualizations and predictions (e.g., "in your opinion, how does the app generate the predictions?"); (2) their willingness to use a fertility app for conception or contraception (e.g., "how likely would you be to use this or a similar app as your only birth control method?", "how likely would you be to use this or a similar app to try to conceive a child?"); (3) their willingness to track different health indicators if they are or are not used for predictions (e.g., "if you learn that the app uses only your period dates to predict your fertile window how likely would you be to track temperature?”); (4) their trust in the predictions for conception and contraception (e.g., “if you are trying to avoid conception, and your fertile window starts on the 12th, how likely would you be to have intercourse without another form of contraception on the 11th?”). We will also include questions to analyze participants’ general fertility knowledge, use of fertility apps, attitudes towards technology, and demographics. We hypothesize that AI descriptions would positively influence people’s willingness to use fertility apps and their trust in the results. We also hypothesize that participants would be more willing to track health indicators that are used by the apps to generate fertility predictions. The session is expected to take one hour. We plan to recruit around 50 female participants of varying expertise in fertility and fertility self-tracking.

Contributions

This study will help us examine to what extent people understand fertility algorithms and trust their results, especially considering the lack of information provided by most currently available fertility apps. As AI becomes increasingly pervasive in our everyday life, it is imperative to understand how lay persons understand, trust, and use algorithmic systems in making important personal decisions, and the broader, societal impact of their widespread adoption.

References

1. Alper T. Alan, Mike Shann, Enrico Costanza, Sarvapali D. Ramchurn, and Sven Seuken. 2016. It is Too Hot: An In-Situ Study of Three Designs for Heating. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, ACM, 5262–5273.
2. Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, et al. 2019. Guidelines for Human-AI Interaction. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, ACM, 3:1–3:13.
3. Adejoke B. Ayoola, Gail L. Zandee, and Yenupini J. Adams. 2016. Women’s Knowledge of Ovulation, the Menstrual Cycle, and Its Associated Reproductive Changes. Birth 43, 3: 255–262.
4. 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: 385–397.
5. C. D. Wickens, J. G. Hollands, S. Banbury, and R. Parasuraman. 2013. Engineering psychology and human performance. Pearson.
6. Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, et al. 2019. Explaining Decision-Making Algorithms Through UI: Strategies to Help Non-Expert Stakeholders. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, ACM, 559:1–559:12.
7. Mayara Costa Figueiredo, Clara Caldeira, Elizabeth V. Eikey, Melissa Mazmanian, and Yunan Chen. 2018. Engaging with Health Data: The Interplay Between Self-Tracking Activities and Emotions in Fertility Struggles. Proc. ACM Hum.-Comput. Interact.
8. Marguerite Duane, Alison Contreras, Elizabeth T. Jensen, and Amina White. 2016. The Performance of Fertility Awareness-based Method Apps Marketed to Avoid Pregnancy. Journal of the American Board of Family Medicine: JABFM 29, 4: 508–511.
9. Jordan Eschler, Amanda Menking, Sarah Fox, and Uba Backonja. 2019. Defining Menstrual Literacy With the Aim of Evaluating Mobile Menstrual Tracking Applications. Computers, informatics, nursing: CIN.

10. Alexander Freis, Tanja Freundl-Schütt, Lisa-Maria Wallwiener, et al. 2018. Plausibility of Menstrual Cycle Apps Claiming to Support Conception. Frontiers in Public Health 6.

11. Kerry D. Hampton, Danielle Mazza, and Jennifer M. Newton. 2013. Fertility-awareness knowledge, attitudes, and practices of women seeking fertility assistance. Journal of Advanced Nursing 69, 5: 1076–1084.

12. Victoria Hollis, Alon Pekurovsky, Eunika Wu, and Steve Whittaker. 2018. On Being Told How We Feel: How Algorithmic Sensor Feedback Influences Emotion Perception. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 2, 3: 114:1–114:31.

13. Sarah Johnson, Lorrae Marriott, and Michael Zinaman. 2018. Can apps and calendar methods predict ovulation with accuracy? Current Medical Research and Opinion 34, 9: 1587–1594.

14. 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: 767–774.

15. Aaron J Masino, Mary Catherine Harris, Daniel Forsyth, et al. 2019. Machine learning models for early sepsis recognition in the neonatal intensive care unit using readily available electronic health record data. PloS one 14, 2: e0212665.

16. 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 and Gynecology 127, 6: 1153–1160.

17. Aaron Springer, Victoria Hollis, and Steve Whittaker. 2017. Dice in the Black Box: User Experiences with an Inscrutable Algorithm. 2017 AAAI Spring Symposium Series.

18. Simone Stumpf, Vidya Rajaram, Lida Li, et al. 2009. Interacting Meaningfully with Machine Learning Systems: Three Experiments. Int. J. Hum.-Comput. Stud. 67, 8: 639–662.

19. Alice J Watson, Julia O'Rourke, Kamal Jethwani, et al. 2011. Linking electronic health record-extracted psychosocial data in real-time to risk of readmission for heart failure. Psychosomatics 52, 4: 319–327.

20. 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. Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, ACM, 623–634.