User Reception of Babylon Health's Chatbot

Daniela Azevedo[a], Axel Legay[b] and Suzanne Kieffer[c]
Université Catholique de Louvain, Louvain-la-Neuve, Belgium

Keywords: User Experience, mHealth, Babylon Health, Chatbot, Artificial Intelligence.

Abstract: Over the past decade, renewed interest in artificial intelligence systems prompted a proliferation of human-computer studies studies. These studies uncovered several factors impacting users' appraisal and evaluation of AI systems. One key finding is that users consistently evaluated AI systems performing a given task more harshly than human experts performing the same task. This study aims to uncover another finding: by presenting a mHealth app as either AI or omitting the AI label and asking participants to perform a task, we evaluated whether users still consistently evaluate AI systems more harshly. Moreover, by picking young and well educated participants, we also open new research avenues to be further studied.

1 INTRODUCTION

Upcoming breakthroughs in critical elements of technology will accelerate the development of artificial intelligence (AI) and multiply its application to face global economic, social, and ecologic crises. Understanding AI's capabilities as well as its limits is quintessential to progress towards context-mindful models. Yet, little is known regarding everyday users' reception of this thrilling and intimidating technology. The worsening healthcare crisis deriving from an ageing population coupled with a global pandemic and a decrease in public health funding, urges practitioners and administrators alike to find new solutions to provide affordable care. AI powered systems are not only cheaper in the long run, but also exponentially increase the accuracy and computability required to produce strategies individually tailored to each patient, making precision medicine possible and available to a majority of the population on a day-to-day basis. With the help of AI, healthcare systems can move away from one-size-fits-all types of care that produce ineffective treatment strategies, which sometimes result in untimely deaths (Buch et al., 2018).

Recent years have seen a sharp increase in mHealth applications, both in the form of websites as well as mobile applications. The services offered by mHealth apps range from disease-specific, doctor-prescribed to generalist, occupation-based apps. However, consumers' general reluctance to engage with AI for sensitive subjects makes implementing AI based solutions complicated. User Experience (UX) offers a considerable arsenal of tools to measure the reception of such applications and improve them to increase adoption rates (Lew and Schumacher, 2020). By analysing the reception of the case of chatbot, a particularly popular technology (Cameron et al., 2018), we aim to shed some light on this subject and thusly facilitate further research.

2 PREVIOUS RESEARCH

2.1 AI-based Technologies in Medicine

Artificial Intelligence (AI) is an umbrella term that refers to technologies ranging from machine learning, natural language processing, to robotic process automation (Davenport et al., 2020). The idea behind the term AI is that the program, algorithm, systems and machines demonstrate or exhibit aspects of intelligence and mimic intelligent human behaviour. Several technological breakthroughs are expected to galvanize the AI-researchers community and significantly transform fields in which it is applied (Gruson et al., 2019; Campbell, 2020). In particular, chatbots are predicted to soon become users' preferred interface, over traditional webpages or mobile applications (Cameron et al., 2018). In the United Kingdom, even before the pandemic, over a million users already preferred to use a chatbot app called Babylon – an app

[a] https://orcid.org/0000-0003-0426-3206
[b] https://orcid.org/0000-0003-2287-8925
[c] https://orcid.org/0000-0002-5519-8814

Azevedo, D., Legay, A. and Kieffer, S.
User Reception of Babylon Health's Chatbot.
DOI: 10.5220/0010803000003124
In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2022) - Volume 2: HUCAPP*, pages 134-141
ISBN: 978-989-758-555-5; ISSN: 2184-4321
Copyright © 2022 by SCITEPRESS – Science and Technology Publications, Lda. All rights reserved

that uses a question-based interaction with users regarding their disease symptoms to establish a diagnosis – rather than contacting their National Health Services (NHS) (Chung and Park, 2019).

Despite reservations regarding the implementation of AI emerging from both users and healthcare professionals (Zeitoun and Ravaud, 2019; Lew and Schumacher, 2020), healthcare slowly departs from an expert-led approach towards a patient-centred, independent and self-sufficient one. Since the democratisation of the internet, patients are increasingly looking to (re)gain control over their healthcare choices (Dua et al., 2014) and the number of people turning to websites and social media in search of healthcare information is only increasing (Chowriappa et al., 2014). AI-systems could offer a more suitable alternative to expert-led healthcare, as opposed to the raw healthcare information currently available on the internet. Using AI-enabled applications allows users to become more active participants in their own health, reduce costs, improve patient experience, physician experience and the health of populations (Campbell, 2020).

User resistance increases as AI moves towards context awareness (Davenport et al., 2020). Therefore, for AI be successfully implemented, lay users' perceptions and beliefs need to evolve. Researchers in the early 2000s were already calling for greater attention to consumer resistance regarding technology alternatives (Edison and Geissler, 2003) – the more reservations users emit, the less likely they are to adopt AI-based products. Somat (Distler et al., 2018) describes the process of implementing and adopting a new product as a journey through an acceptability-acceptation-appropriation continuum. The process begins with a subjective evaluation before use (acceptability), after use (acceptation) and once the product has become a part of daily life (appropriation).

Holmes et al. (2019) argue that interacting with a chatbot is more natural and more intuitive than conventional methods for human-computer interactions. A chatbot is an intelligent interactive platform that enables users to interact with AI through a chatting interface (Chung and Park, 2019) using natural language (written or spoken) aiming to simulate human conversation (Denecke and Warren, 2020). De Gennaro et al. (2020, p. 3) note that 'chatbots with more humanlike appearance make conversations feel more natural, facilitate building rapport and social connection, as well as increase perceptions of trustworthiness, familiarity, and intelligence, besides being rated more positively'. To decrease complexity, most chatbots opt to restrict user input to selectable predefined items (Denecke and Warren, 2020). There are already chatbots in the market that provide therapeutics or counselling, disease or medication management, screening or medical history collection or even symptoms collection for triage purposes (Denecke and Warren, 2020). These AI applications can instantly reach large amounts of users (de Gennaro et al., 2020), reduce the need of medical staff, as well as assist patients regardless of time and space (Chung and Park, 2019).

2.2 UX's Role in AI Adoption

UX plays a pivotal role in the acceptance process (Lew and Schumacher, 2020). UX is a compound of emotions and perceptions of instrumental and non-instrumental qualities that arise from users' interaction with a technical device. To achieve a successful UX, AI products, like any other product, need to meet essential elements of utility, usability and aesthetics (Lew and Schumacher, 2020). Wide-spread consumer adoption requires good usability (Campbell, 2020), which is currently lacking and thus undermining efforts to deliver integrated patient-centred care.

UX is composed of many constructs, which are complex and sometimes impossible to measure holistically (Law et al., 2014). For example, trust, which can be defined as 'the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability' (Lee and See, 2004, p. 54), requires a balanced calibration to be struck, especially regarding health-related issues. Overtrust results in a users' trust exceeding system capabilities, while distrust leads users to not fully take advantage of the system's capabilities. Two fundamental elements define the basis for trust: the focus of what is to be trusted and the information regarding what is to be trusted.

Low adoption of AI-systems is linked to low trust rates (Sharan and Romano, 2020), because trust strongly influences perceived success (Lew and Schumacher, 2020). One of the main reasons behind AI's failure to build user trust is that their output is often off-point, thus lacking in accuracy. Trust is primarily linked to UX pragmatic qualities, in particular usability and utility. Aesthetics influences the perception of both usability and utility through a phenomenon called the aesthetics usability effect. This refers to users' tendency to perceive the better-looking product as more usable, even when they have the exact same functions and controls (i.e., same utility and usability) (Norman, 2005).

2.3 UX Challenges

Davenport et al. (2020) distinguishes four main challenges for AI adoption: (1) users hold AI to a higher standard and are thus less tolerant to error; (2), users tend to be less willing to use AI applications for tasks involving subjectivity, intuition or affect, because they believe having those characteristics is necessary to successfully complete the task ; (3) users are more reluctant to use AI for consequential tasks, such as driving a car as opposed to choosing a movie, because it involves a higher risk and (4) user characteristics (e.g., gender) also influence perception, evaluation and adoption of AI. When perceived as a risk, women tend to adopt AI less since they are more risk adverse than men (Gustafson, 1998). Individual's attitude towards technologies is another key factor (Edison and Geissler, 2003). Furthermore, having little technical or digital literacy ultimately influences users' perception of it.

According to Araujo et al. (2020), level of education and programming knowledge influence perception of AI: people with a lower level of education show a stronger negative attitude towards algorithmic recommendations and participants with higher levels of programming knowledge show a higher perceived fairness level. In general, people with domain-specific knowledge, belief in equality and online self-efficacy tend to have more positive attitudes about the usefulness, the fairness and risk of decisions made by AI and evaluate those decisions on par and sometimes better than human experts. However, when users strongly identify with the domain activity, they are less inclined to adopt AI for that activity (Davenport et al., 2020).

Another concern is uniqueness neglect. In a study carried out by Longoni and al. (2019), users have reservations about AI because they generally believe that AI only operates in a standardized manner and is calibrated for the 'average' person. Consequently, AI-systems are perceived as less able to identify and account for users' unique characteristics, circumstances and symptoms, leading them to be neglected. Yet, because same diseases show very different symptoms based on patient's health condition or lifestyle (Chung and Park, 2019), it is human doctors who too often misdiagnose patients. AI is able to simultaneously take into account patient's medical records and family history, and even their genome, warn about the disease risk and design unique treatment pathways tailored for them (Panesar, 2019). AI-systems could help prevent misdiagnoses by presenting doctors with alternative diagnoses and information and thusly supporting and enhancing them.

3 HYPOTHESES

H_1. Men and women evaluate products differently. When perceived as a risk, women tend to adopt AI less since they are more risk adverse than men (Gustafson, 1998). Since then, gender dynamics have evolved. Therefore, women today may no longer tend to evaluate AI – a technology perceived to involve more risk – significantly differently than their male counterparts.

H_2. Products' hedonic (Attractiveness, Novelty) and pragmatic qualities (Content Quality, Trustworthiness of Content, Efficiency, Perspicuity, Dependability, Usefulness) are evaluated more negatively when the product is explicitly labelled AI. Merely labelling a product as AI may trigger in laypeople a variety of heuristics based on stereotypes about the operation of the application (Sundar, 2020). The outcome and the extent of these triggers on subjective evaluation is yet to be determined.

H_3. User characteristics (attitude towards technology, English proficiency, domain-specific knowledge) interacts with the treatment and significantly changes their subjective evaluation of the product. User characteristics is an overarching factor of subjective evaluation. However, we do not know if users' evaluation differs significantly depending on the nature of the product. For example, users' attitude towards technology underlines the overall reception and resistances towards technologies (Edison and Geissler, 2003); language proficiency plays a pivotal role as it can hinder users' ability to understand the system; users' domain-specific knowledge (e.g., programming knowledge or products' domain of activity) influences the evaluation of related AI products (Araujo et al., 2020).

4 METHOD

4.1 Why Babylon Health?

We chose Babylon Health, a free-to-use webapp that provides various healthcare services. We focused on their Chatbot Symptom Checker. According to Babylon Health's website, their chatbot is powered by an AI that understands symptoms the users input and provides them with relevant health and triage information. Chung and Parker (2019) confirm that users receive responses depending on their input, which are based on data contained in a large disease database. After the symptom check, users receive a diagnosis and a suggested path of action (e.g., appointment with GP). Since the scientific literature suggests trust

in AI-systems often correlates positively with higher degrees of anthropomorphism (Sharan and Romano, 2020), we deliberately chose an AI-system that has no human-like element (behavioural or visual), to avoid biasing the results with features which are more prone to positive reception.

4.2 Participants

Typically, the last population segment to adopt innovations are older people and people with lesser educational attainment and lower socioeconomic status (Dorsey and Topol, 2020). Therefore, to avoid biasing the results by obtaining a sample with too many variables, we recruited exclusively young, well-educated participants. We did not accept participants who had previously engaged with either Babylon or very similar applications, such as Ada Health, because past experiences with AI-systems may significantly alter the trust formation process (Hoff and Bashir, 2015).

4.3 Data to Collect

Regarding gender (H_1), participants self-identified as either (1) male, (2) female or (3) non-binary.

Regarding pragmatic and hedonic constructs (H_2), we adopted a mixed approach (Law et al., 2014). As quantitative method, we used the UEQ+ (Schrepp and Thomaschewski, 2019), a questionnaire intended for evaluating UX through the lens of up to 26 UX constructs. Since Babylon qualifies both as a word processing device and an information website, we selected the UX constructs of dependability, efficiency, perspicuity, content quality and trustworthiness of content. Since perceived usefulness is paramount to adoption of AI (Araujo et al., 2020), we include a usefulness scale. Furthermore, to measure at least two hedonic qualities, we added the well-rounded staple of UX that is attractiveness scale and, because the product is based on AI, a novelty scale. Since perspicuity and dependability are jargon words, we provided a definition. We also asked participants to rank the eight UX constructs to establish a hierarchical scale. As qualitative method, we utilised the interview to ask participants to explain their ranking in their mother language (French).

Regarding user characteristics (H_3), we administered a questionnaire to assess attitude towards technology found in Edison and Geissler (2003). The questionnaire measures technophobia with a 10 statements agreement scale. To facilitate user input, we continued using a 7-point Likert-scale instead of the proposed 5. We calculated 'measured technophobia' as follows: mean scores between 1 and 2.49 are considered *Highly Technophobic*, from 2.5 to 3.99 *Moderately Technophobic*, from 4 to 5.49 *Mildly Technophobic* and from 5.5 to 7 *Not Technophobic*. To determine if participants were laypeople, we asked them if they either study, work or are interested in either health or IT. This corresponds to their domain-specific knowledge. We evaluated participants' English proficiency by assessing the nature and frequency of their vocabulary questions. We also took into account participants' self-reported proficiency.

4.4 Method of Collection

We conducted two rounds experiment. Due to COVID-19, experiment 1 (XP1) took place on Zoom, a cloud-based video communication app that allows virtual video and audio set up, screen-sharing and recording. Participants were at home and used their personal computer. Experiment 2 (XP2) took place on university campus, and a computer was put at the participants' disposal. The computer recorded the screen, webcam and sound. In both experiments, we used LimeSurvey to conduct the survey and to give the instructions.

We used a randomized, post-test-only experimental design to test our hypotheses since these designs are well suited to detect differences and cause-effect relationships. Since today most people are unwilling to take a 30-minute survey and participants' attention and accuracy declines over time (Lew and Schumacher, 2020; Smyth, 2017), we decreased question and task complexity as the experiment progressed. We broke down the experiment into four parts: (1) a user test (10 minutes), (2) a UX questionnaire and ranking of UX constructs (12 minutes), (3) a short interview (4 minutes) and (4) an attitude towards technologies test followed-up by a sociodemographic questionnaire (5 minutes).

To understand the influence of the term 'Artificial Intelligence' on users' experience, we divided our participants into two groups. Group 1 (*No Treatment Group*, also known as control group) received an introduction of the product that did not mention AI. Group 2 (*Treatment Group*) received the almost same introduction, except we mentioned AI 14 times : twice explicitly when we presented the experiment orally and when we introduced the app in the written instructions, and then 12 times implicitly by reminding them of this element in the title of the survey and the subtitles above each section of the survey. These were the sole differences in treatment between the two groups.

4.5 Procedure

First, we briefly explained how the experiment would unfold. Then, we asked participants to share their screen (XP1) or gave them the survey directly on the computer, with the recording activated and their screen duplicated on a monitor (XP2). Participants then read a brief description of Babylon Health. They proceeded to read a situation in which they were asked to use Babylon Health to find a diagnosis. We instructed them to use the product according to the symptoms they were given. We told them that if a symptom or an act was not mentioned, then it had not happened in that situation. Participants were allowed to read the situation as many times as they wished, while they were using the product and inputting their symptoms. They were instructed to return to the survey once they had reached the result page and had read the results to their liking.

Then, we asked them to share their impressions and subjectively evaluate the product by completing the UEQ+. Next, we asked participants to rank UX constructs by order of importance. We followed-up on their ranking answers with a semi-structured interview conducted to obtain a more precise account of their thoughts and impressions. We asked participants (1) to explain the reasoning behind their ranking, (2) to describe their experience, (3) whether the product met their expectations, (4) if something worked differently than they expected, and (5) after stop sharing their screen to answer questions related to their attitude towards technology and to rate their level of technophobia.

We then asked participants to tell us if they noticed Babylon was made of AI and if so, to disclose the moment they knew the app was AI-enabled and to discuss the impact of this new information had on their overall impression of the product. They could either have a more negative opinion, have a more positive opinion or not change their impression at all. We only report the answers of the No Treatment group, since the Treatment group knew it all along.

Finally, we ask them to disclose their knowledge of Babylon Health to see if participants had previous knowledge of the app and see if participants in the Treatment Group had noticed Babylon was AI-enabled. We also ask them to input their sociodemographic information (age, gender, highest degree of education, current employment status, field of work/study and interests), as this information is closely related to other social constructs pertaining to users' characteristics, which therefore influences their user experience.

5 RESULTS

5.1 Experiment 1

Experiment 1 took place remotely between February 24th and April 18th 2021 and involved 29 participants (10 males), aged from 21 to 34 (mean 24.1), among which 21 (72%) were enrolled bachelors and master students, while eight (28%) had already obtained a master's degree, mostly in communication studies (69%). Twelve out of 29 participants (41%) barely had sufficient mastery of English to follow the instructions and interact with the application. Further, ten participants strongly deviated from the expected path in the application, by inputting symptoms at the beginning of the interaction, which were either very broad or incorrect, and led to unrelated questions and results from Babylon.

Participants receiving no treatment overwhelmingly (80%) reported that knowing the app is made of AI did not influence their subjective evaluation of the product, while two participants (13%) reported having a better impression of the product after receiving this information and one (7%) perceived it to be worse. Regardless of treatment, participants evaluated the product positively ($mean \geq 5.3$ out of 7). Perspicuity and Content Quality got a particularly positive mean score (≥ 6), while Novelty got a mean score of 4.85. Constructs scores varied from a minimum of 2.25 points (Content Quality), to a maximum of 4.75 points (Novelty), which indicates the presence of strong outliers.

H_1. Data showed no statistical difference between male and female. We used a Student's t-test when groups were normally distributed, a Mann-Whitney test when they were not. Further, the Multivariate Test showed no significant result for the Treatment x Gender interaction, $t(8,18)=.585$, $p=.777$, partial $\eta^2=.206$. Given these results, we do not discriminate based on gender.

H_2. Data showed no statistical difference in the evaluation of the product based on treatment (Mann-Whitney U-test, Student's t-test, Multivariate test). However, five constructs (Attractiveness, Perspicuity, Usefulness, Novelty, Trustworthiness of Content had small effect size ($d \leq 0.2$) and three (Efficiency, Dependability, Content Quality) only had a medium effect size ($d \leq 0.5$).

H_3. A Multivariate Test indicated no significant interaction between treatment and participant's characteristics (English language skills, Domain-specific Knowledge, Self-reported Technophobia, Measured Technophobia). Overall, we obtained a large effect size for all interactions (partial η^2 was always $\geq .14$).

5.2 Experiment 2

Experiment 2 took place on campus between August 24 and 31st 2021 and involved 11 PhD students (10 females) from Université catholique de Louvain, aged from 25 to 35 (mean 28). Three participants had an interest for either Health or IT. Participants' English proficiency was always deemed sufficient and only three encountered some difficulty. They all took appropriate paths in the application. All participants in No Treatment Group reported believing that knowing the product is AI-enabled did not influence their subjective evaluation of the product. Participants evaluated the product positively, with seven out of the eight UX constructs reaching a *mean* $\geq$ 5.27 (only Novelty obtained a mean of 4.93).

H_1 + H_3. There were insufficient participants to properly categorise them by their characteristics (gender, English proficiency, domain-related knowledge, attitude towards technologies).

H_2. We were, however, able to measure if participants evaluated the product differently according to treatment. A Student's t-test showed that participants in Treatment group (M=5.650, SD=0.57) evaluated Content Quality more negatively than in No Treatment Group (M=6.33, SD=0.26), $t(9)$=2.629, p=.027, d=1.53. Other constructs had neither significant results nor large size effect.

5.3 Experiment 1 + 2

Participants from XP1 and XP2 ranked the level of importance of UX constructs similarly (Table 1). Trustworthiness of Content and Content Quality continue to be designated as the most important qualities for an Health application, whereas Attractiveness and Novelty are cited as the least important. Interviews showed that Content Quality is regarded as one of the most important qualities more frequently than written responses suggest. When asked to explain their ranking, participants focused on aspects linked to Trustworthiness and Content Quality. Trust was a word that came up particularly often and was treated at length. Aspects related to Content Quality and Dependability often intertwined with this 'trust' concept. A Student t-test showed no statistical difference between XP1 and XP2. Furthermore, a Levene's test indicated that the variances for all constructs in the two groups were equal. Therefore, we were able to combine the results from the two experiments to gain more statistical power.

H_1. We found no significant difference between genders. However, since women outweighed men 3 to 1, this analysis bears no scientific value.

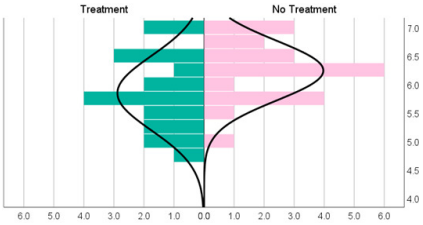

Figure 1: Population Pyramid Frequency with Bell Curve of Content Quality by Treatment, positive scale (4 to 7).

H_2. Data showed a significant difference between treatment groups in their evaluation of Babylon's Content Quality ($t(38)$=2.111, p=.041, d=.664). Especially, the No Treatment Group (N=21, M=6.238, SD=.527) scored Content Quality higher than the Treatment Group (N=19, M=5.842, SD=.657). This means participants in the No Treatment Group found the app to be more up-to-date, interesting, well prepared and comprehensible. No other construct obtained a significant result, but they also have a very inferior effect size. Usefulness, Novelty, Trustworthiness of Content had a very small effect size ($d \leq 0.1$) Attractiveness, Efficiency, Perspicuity had a small effect size ($d \leq 0.2$), and Dependability had a medium effect size ($d \leq 0.5$).

H_3. User characteristics (English proficiency, domain-specific knowledge, attitude towards technology) did not significantly interact with the treatment.

6 DISCUSSION

H_1. Our study results indicate female and male participants assess their UX with Babylon similarly, which contradicts previous work on the importance of gender in subjective evaluation of AI. Women should be more risk-adverse than men (Gustafson, 1998). This may indicate that gender related association to risk might be outdated and needs to be revisited. Differences between genders might be eroding in millennials and Gen Z. Another explanation is that millennials and Gen Z do not perceive AI as a risk. This is plausible if we consider that participants are not entering real symptoms to find a real diagnosis to a real illness. Rather, they are entertaining a fictional scenario simply because we asked them to do so. Thus, they might not perceive the use of Babylon as an increase in risk. The outcome being imaginary and not translated into actual consequences might simply cancel the risk perception. If so, then participants may also not feel concerned by uniqueness neglect (Longoni et al., 2019), because they only perceive themselves as unique and not others, even when those 'others' are their hypothetical sick selves.

Table 1: Construct ranked by importance. We calculated scores by weighing the position of each construct in the ranking with points (1^{st} = 8 pts, 2^{nd} = 7 pts...) and the Mean Rank from their average points.

Rank	Constructs	Score	Mean Rank	Most important	Least Important
1	Trustworthiness of Content	197	6.64	33	0
2	Dependability	167	5.70	9	0
3	Usefulness	159	5.25	8	0
4	Perspicuity	143	4.82	9	0
5	Content Quality	142	4.71	21	0
6	Efficiency	140	4.70	7	2
7	Attractiveness	75	2.35	0	21
8	Novelty	58	2.05	0	25

H_2. We require more participants to see whether labelling a product as 'AI' changes the UX with the product. However, the significant difference regarding Content Quality suggests more research might be of interest regarding. Further, our results comply with previous findings: the most important aspect in UX is trust, a concept without which participants said they would never adopt an app; users hold AI systems to a higher standard than other applications; and users evaluate AI systems that fail more harshly.

H_3. The lack of interaction with the treatment might indicate that user characteristics influence their reception of products, regardless of the 'AI' label. The non-significant interaction between treatment and participants' attitude towards technology may be due to participants' profile. Young and well-educated people are not typically resistant to adopting new technologies, and might have equal reservations towards both AI and non-AI-systems regarding health. As for domain-specific knowledge, the mere interest in a subject is not enough to significantly alter users' subjective evaluation of a product. Thus, assuming previous research findings apply to the younger generation, participants need to either study or work in the related domain to alter their evaluation of AI. Lastly, language skills did not significantly interact with treatment, suggesting other key factors are at the origin of difference in results between treatments.

Limits and Future Research. Further research involving a larger number of participants with similar characteristics should be conducted to fully answer our hypotheses, since small size effect prevents us from drawing definitive conclusions. About 100 participants per treatment group should be recruited to acquire a meaningful size effect with a Confidence Interval of 95% and normal distribution. To be representative and useful to practitioners, this experiment should also be replicated with less educated or older populations and extended to different apps. In addition, to achieve action fidelity (Kieffer, 2017), similar research should be conducted on users who are genuinely sick with the same disease at the time of the experiment to avoid artificial conditions and decrease error rate related to symptom input. The product would likely seem more intuitive, since users would not have to follow a scenario. Reservations regarding effects such as users' uniqueness neglect concerns, if applicable, would then manifest themselves.

7 CONCLUSION

Our results manage to contribute to research in two ways. First, Gustafson's theory on gender biases as pertaining to risk assessment might need to be revisited. Differences between genders are eroding and becoming less relevant and appropriate to use when dealing with younger generations. They may also be much more context-dependent than previously thought: a hot topic such as health during a pandemic may soften or override gender-based variation in perception to such an extent it might even become irrelevant. Second, though XP1 was conducted remotely, we were still able to obtain actionable data: XP1 and XP2 showed very similar results, suggesting that conducting UX experiments out of the lab is possible without inconsolably damaging the results. They may, however, require more participants to reach the same conclusions. This finding should be celebrated, as we are likely to enter an era of pandemics that will force us to change how we conduct experiments and continue to produce valid and scientific work.

Healthcare systems require a momentous the help of AI to properly undertake current and future challenges, such as an increase in age-related illnesses due to ageing populations. Detecting the differences in behaviour, perception and treatment remains an important subject to explore, since this information enables UX practitioners to design counter strategies tailored for medical AI apps' unique needs.

UX research would benefit from clear, flexible and extensive frameworks that can be used in a plethora of contexts to evaluate users' perceptions. Moreover, these need to be widely employed because gaining

insights from cross-study comparisons and thus having large bodies of comparable works is quintessential to properly interpret components that influence users' subjective evaluation.

REFERENCES

Araujo, T., Helberger, N., Kruikemeier, S., and de Vreese, C. H. (2020). In AI we trust? perceptions about automated decision-making by artificial intelligence. *AI & SOCIETY*, 35(3):611–623.

Buch, V. H., Ahmed, I., and Maruthappu, M. (2018). Artificial intelligence in medicine: current trends and future possibilities. *British Journal of General Practice*, 68(668):143–144.

Cameron, G., Cameron, D. W., Megaw, G., Bond, R. R., Mulvenna, M., O'Neill, S. B., Armour, C., and McTear, M. (2018). Best practices for designing chatbots in mental healthcare – a case study on iHelpr. In *Proceedings of the 32nd International BCS Human Computer Interaction Conference*.

Campbell, J. L. (2020). Healthcare experience design: A conceptual and methodological framework for understanding the effects of usability on the access, delivery, and receipt of healthcare. *Knowledge Management & E-Learning*, 12(4):505–520.

Chowriappa, P., Dua, S., and Todorov, Y. (2014). Introduction to machine learning in healthcare informatics. In Dua, S., Acharya, U. R., and Dua, P., editors, *Machine Learning in Healthcare Informatics*, pages 1–23. Springer Berlin Heidelberg.

Chung, K. and Park, R. C. (2019). Chatbot-based heathcare service with a knowledge base for cloud computing. *Cluster Computing*, 22:1925–1937.

Davenport, T., Guha, A., Grewal, D., and Bressgott, T. (2020). How artificial intelligence will change the future of marketing. *Journal of the Academy of Marketing Science*, 48(1):24–42.

de Gennaro, M., Krumhuber, E. G., and Lucas, G. (2020). Effectiveness of an empathic chatbot in combating adverse effects of social exclusion on mood. *Frontiers in Psychology*, 10:3061.

Denecke, K. and Warren, J. (2020). How to evaluate health applications with conversational user interface? *Studies in Health Technology and Informatics*, 270:970–980.

Distler, V., Lallemand, C., and Bellet, T. (2018). Acceptability and acceptance of autonomous mobility on demand: The impact of an immersive experience. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 1—-10, New York, NY, USA. Association for Computing Machinery.

Dorsey, E. R. and Topol, E. J. (2020). Telemedicine 2020 and the next decade. *The Lancet*, 395(10227):859.

Dua, S., Acharya, U. R., and Dua, P., editors (2014). *Machine Learning in Healthcare Informatics*, volume 56 of *Intelligent Systems Reference Library*. Springer Berlin Heidelberg.

Edison, S. and Geissler, G. (2003). Measuring attitudes towards general technology: Antecedents, hypotheses and scale development. *Journal of Targeting, Measurement and Analysis for Marketing*, 12:137–156.

Gruson, D., Helleputte, T., Rousseau, P., and Gruson, D. (2019). Data science, artificial intelligence, and machine learning: Opportunities for laboratory medicine and the value of positive regulation. *Clinical Biochemistry*, 69:1–7.

Gustafson, P. E. (1998). Gender differences in risk perception: Theoretical and methodological erspectives. *Risk Analysis*, 18(6):805–811.

Hoff, K. A. and Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 57(3):407–434. PMID: 25875432.

Kieffer, S. (2017). Ecoval: Ecological validity of cues and representative design in user experience evaluations. *AIS Transactions on Human-Computer Interaction*, 9(2):149–172.

Law, E. L.-C., van Schaik, P., and Roto, V. (2014). Attitudes towards user experience (UX) measurement. *International Journal of Human-Computer Studies*, 72(6):526–541.

Lee, J. D. and See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, page 31.

Lew, G. and Schumacher, R. M. (2020). *AI and UX: Why Artificial Intelligence Needs User Experience*. Apress.

Longoni, C., Bonezzi, A., and Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research*, 46(4):629–650.

Norman, D. A. (2005). *Emotional Design: Why We Love (or Hate) Everyday Things*. Basic Books, 3rd edition.

Panesar, A. (2019). *Machine Learning and AI for Healthcare: Big Data for Improved Health Outcomes*. Apress, 1st edition.

Schrepp, M. and Thomaschewski, J. (2019). Design and validation of a framework for the creation of user experience questionnaires. *International Journal of Interactive Multimedia and Artificial Intelligence*, 5(7):88.

Sharan, N. N. and Romano, D. M. (2020). The effects of personality and locus of control on trust in humans versus artificial intelligence. *Heliyon*, 6(8):e04572.

Smyth, J. D. (2017). *The SAGE Handbook of Survey Methodology*, pages 218–235. SAGE Publications Ltd.

Sundar, S. S. (2020). Rise of Machine Agency: A Framework for Studying the Psychology of Human–AI Interaction (HAII). *Journal of Computer-Mediated Communication*, 25(1):74–88.

Zeitoun, J.-D. and Ravaud, P. (2019). L'intelligence artificielle et le métier de médecin. *Les Tribunes de la santé*, 60(2):31–35.