A Replication Study for IoT Privacy Preferences

Ahmed Alhazmi, Ghassen Kilani, William Allen and TJ OConnor
Computer Engineering and Sciences
Florida Institute of Technology
Melbourne, United States of America
aalhazmi2017@my.fit.edu, gkilani2010@my.fit.edu, wallen@fit.edu, toconnor@fit.edu

***Abstract*—Privacy issues have plagued the rapid proliferation of the Internet of Things (IoT) devices. Resource-constrained IoT devices often obscure transparency for end-users. The usability of privacy-preserving mechanisms and tools offers promise but relies on accurately capturing privacy preferences. Our work replicates a previous study to examine users' privacy expectations and preferences for IoT devices. We specifically focus our effort on examining users' feelings regarding their data collection in an IoT-based environment. Our work analyzes different contributing factors that impact users' privacy decisions about data collection. Our analysis supports previous work that has argued users' perceived benefit is an essential factor and motivating favor. In contrast to the previous study, we identified the workplace has now morphed into a sensitive location where users are uncomfortable sharing their private information.**

***Keywords*—Privacy, IoT, social aspects, preferences, access control**

I. Introduction

The Internet of Things has been one of the most rapidly evolving technologies in the past few years [1]. This emerging technology facilitated access to electronic services such as health care, assisted living and driving, security, and social interaction while creating new security issues and breaches never dealt with before. The fragile nature of the IoT security and its scalability make it the perfect target for malicious activity and privacy leaks [2]. IoT security has evolved over the recent years to enhance data privacy and integrity owing to its gradual evolution towards becoming a considerable driver in controlling and monitoring applications [3]. The billions of things in the network are powered using the user's private information to provide personalized services and tailored experiences. However, many users are generally unaware of the sensitive information being gathered about them nor understand what this private data will be used for outside of the network [4].

In order to thoroughly achieve the potential of IoT Technology, individuals should be fully aware of sensor interactions and data captured during communications to make knowledgeable decisions. To do so, smart devices should always inform their users about captured private information and respect their preferences and decisions. Solving these problems requires a complex understanding of users' social norms, culture, context, and education as privacy needs are perceived differently between individuals [5]. This dilemma worsen when multiple devices communicate together and make unpredictable interactions. The first step in resolving this issue is trying to understand and identify the contribution of different factors that might impact users' choices in an IoT environment.

Contributions: Our work reproduces a study of privacy preferences by Naeini et al. [6] that analyzed participants' preferences regarding data collection in various situations. The study examined different scenarios, each constructed from eight factors. Our work replicates this approach and compares the results with the original study. Ultimately, this paper seeks to answer the following questions regarding privacy preferences:

RQ1 : What are the privacy-related preferences of users when their data is collected in an IoT-based environment?

RQ2 : How can the findings of this replicated study be comparable to the findings of the original one?

The remainder of this paper is organized as follows: Section II discusses related work. Section III describes the study's methodology. Section IV presents the obtained results. Section V discusses the observations that we made. Section VI explains limitations and future work. Section VII concludes the paper.

II. Related Work

To protect users from privacy invasions in a climate where sensors are omnipresent, Konings and Schaub [7] developed" PriPref", a privacy context awareness application. Users can broadcast their privacy preferences in the surrounding environment in different scenarios for better coexistence through their application. To better assess user's privacy concerns, the authors first surveyed user's disturbances and strategies used to handle them. Based on the survey results, Konings et al. [7] developed an application that displays the prevalent privacy preferences in neighboring environments. PriPref processes all users' preferences and displays the current environment status to all participants for better privacy effectiveness. Users can also enable dynamic privacy adaptation, where the app will automatically modify users' phone setting to accommodate the current environment preferences. This approach sheds light on an essential aspect of user privacy preferences, where the location plays an important part in users' privacy decisions. Tsai et al. [8] also developed a privacy mobile manager based on contextual privacy preferences called" TurtleGuard." This

978-1-6654-3156-9/21/$31.00 ©2021 IEEE

novel privacy permission manager used a machine-learning algorithm to lessen the decision burden on its users. By selecting few contextual circumstances preferences, the applications provided their users with the necessary feedback from various applications with the ability to modify or audit the automated decisions. The research mentioned above stresses the need for a tailored privacy preference experience unique to each individual and reflect his/her needs in different scenarios.

Prior work in the area of privacy [9]–[11] suggest that capturing users' privacy preferences through privacy profiles, each contains a set of privacy-related configurations/tasks, is an effective solution for capturing privacy preferences accurately. For instance, a few privacy profiles were developed in [10] and used by the proposed privacy assistant. New users will respond to a set of questions to determine which profile is suitable for which user. In other words, the assignment of each profile to each user is dependent on the answers provided by the user. Based on the assigned profile, a number of privacy recommendations are given to the user, which will have the ability to accept or deny these recommendations. According to the researchers, approximately 80% of the recommendations were accepted by the participants [10]. Lin et al. [11] also experimented with the feasibility of categorizing users' privacy preferences through profiles to manage mobile app permissions. Their result indicates that it was possible to cluster users into four different privacy groups that served as a basis for more complex privacy needs. Their findings demonstrate that privacy profiles are a viable technique on mobile apps but may not necessarily reflect users' privacy preferences in an IoT environment where various heterogeneous sensors are deployed.

Lee and Kobsa [12] conducted a survey of 172 participants in an attempt to comprehend the privacy preferences of users in an IoT-based environment. The researchers were able to collect 33,090 responses that were then analyzed and used to cluster participants based on the similarity of their privacy preferences. During the clustering process, the impact of contextual factors, such as the data collector's identity, was examined. Thus, four clusters that are tied to the potential privacy risks of the given IoT scenarios in the survey were built. Consequently, a machine learning model that focuses on predicting the privacy-related decisions of users was constructed. The model was able to predict 77% of users' privacy decisions on either allowing or denying the given IoT scenario. Moreover, This study is similar in its objectives to the original study that we are replicating [6], [12]. However, The decision to replicate Naeini et al. [6] research stems from the fact that they counted for more aspects than previous research. Their work focused on a user-centric approach, realistic and futuristic scenarios, and constantly changing human behaviors. They also took into account the interactions between multiple factors, more than any prior research.

III. Methodology

We leveraged the convenience of Google Forms to craft scenarios and record answers. The totality of the responses were kept on Google servers until the survey was completed and moved on the university facilities when the research was concluded. Additionally, we deleted all Gmail accounts for security and privacy purposes. After gathering all the data, we used "Notion.so" [13] to clean and categorize it by factor of interest. We did not use any scripts for data segmentation as the results were manageable, and we wanted to make sure that all data input was reliable and legitimate. We also used different data scale coding than the original study to make results more significant and understandable. The description of our data scale coding will be discussed in section III-B.

TABLE I
Demographic description of participants.

Gender		Age		Education		Income		IUIPC	
Male	70%	Range	18-70	High School	23.4%	<$15k	29.9%	*Control Factor*	
Female	27%	Mean	29.3	Associate	5.2%	$15k-34k	16.9%	Range	1-7
No Answer	3%			Bachelors	27.3%	$35k-74k	16.9%	Mean [SD]	6.12[1.36]
				Masters/PhD	40.3%	$75k-149k	9.1%	*Awareness Factor*	
				No Answer	3.9%	$150k-199k	0.0%	Range	1-7
						> $200k	1.3%	Mean [SD]	6.35[1.17]
						No Answer	26.0%	*Collection Factor*	
								Range	1-7
								Mean [SD]	6.19[1.14]

Recruitment: We started the recruitment process after we received the approval for conducting this study from our university's Institutional Review Board (IRB) on January 25, 2020. All of our participants in the study were recruited using convenience sampling methods (e.g., emails to mailing lists, personal invitations, word of mouth) and snowball sampling methods (participants tell others who then participate). However, unlike the original study, our participants were not required to be residing in the United States. To limit the expected responses to one response per person, each participant in the user study had to use their Gmail email account to access Google Forms and complete the survey.

Participants: Our survey included 77 participants, predominately affiliated with our university. Table I depicts the demographic information of our participants.

A. Design:

Similar to the original survey in [6], we chose to employ a vignette study. We created eight realistic scenarios and four different versions to diversify the possible outcome of our vignette study. Each of those 32 scenarios was meticulously crafted to mimic real-world examples in a specific IoT environment. Seven factors were used to create different scenarios. These factors are (location, data type, device type, user benefit, purpose, retention, shared). Location refers to the place at which the data is collected. Data type refers to the type of data that is collected. Device type refers to the device that is utilized to collect the data. User benefit refers to the entity that benefits from this data collection (it can be the collector of this data or the person whose data is collected or both). Purpose refers to the goal of this data collection. Retention refers to how long this collected data will be kept. Shared refers to the possibility of sharing the collected data. Unlike the original study in [6], we didn't include a factor that focuses on the possibility of inferring information from the collected data in our design process.

In addition, each factor has different levels. For instance, the retention factor has five levels. These levels are forever, week, year, until the purpose is satisfied and unspecified. Integrating these factors will allow us to simultaneously study their impact and importance in participants' decisions regarding privacy. To make our scenarios reliable, we attempted to introduce these factors in the same order as not to confuse the participants. All factors and their levels are presented in table II. It is worth noting that we made some changes to the levels of data type compared to the original study. We grouped the original study's presence and specific position factors and labeled them as presence in our study. We also added temperature as a data type level.

The eight scenarios included each level of each factor except for the purpose and device type factors. These factors are dependent on other factors and so including each level of these two factors might not be always feasible. The following is an example of one the scenarios that were presented to the participants:

- *You are at **home**. Your home has an **iris scanner** that is used to give you access to your home office. The biometric data will be **shared with the device manufacturer** for **security purposes**. **Data retention is unknown**.*

For each scenario, the subjects were given a few questions with a range of possible response choices that they could select from to reflect their perception (user perceived benefit), comfort level, and willingness to allow or deny data collection in that scenario. User perceived benefit is different from user benefit, which was included as a part of our design. User perceived benefit focuses on gathering the perception of participants on each and every scenario that was presented to them. Unlike the original study that didn't consider user perceived benefit for some scenarios in the analysis phase, our analysis process takes into account the user perceived benefit for all scenarios. Naeini et al. [6] excluded user perceived benefit for scenarios that lacked the purpose factor. Subjects were also asked how often they would like to be notified about the collection of their data for each of the represented scenarios. Consequently, a total of three questions was given to the subjects to which they could discuss the factors that affected their comfort level regarding data collection in the represented scenarios and how frequently they would like to see a summary of their collected data. In addition to asking questions tailored to the given scenarios, participants were provided with the same Internet Users' Information Privacy Concerns (IUIPC) statements used in the original study [6]. Subjects could express their agreement level with each of the given statements through a number of possible response choices ranging from "Strongly Agree" to "Strongly Disagree". The subjects' responses to these statements are used to calculate the IUIPC score, which is shown in table I. Moreover, a number of demographic questions were provided to the subjects, such as questions regarding their age and income.

Each participant was required to read and agree to an informed consent form before partaking in the study. After that, the participant would be able to read each scenario and respond to questions about it. Then, subjects would be asked to respond to the rest of the questions. Finally, upon completing the study, the participants had the choice to enter a draw to win 1 of 4 Amazon gift cards. Each Amazon gift card is worth $10. On a final note, our methodology and design process didn't involve designing or building any prediction models as this was not one of the goals of our replication study.

TABLE II
BREAKDOWN OF FACTORS AND THEIR LEVELS.

Factor	Levels
location	department store; library; workplace; friend's house; home; public restroom
data type	presence; temperature; video; biometric (e.g. fingerprint recognition)
purpose	mentioned; not mentioned
shared	mentioned; not mentioned
data retention	week; year; forever; until purpose is satisfied; unspecified
user benefit	user; collector; both
device type	camera; fingerprint scanner, facial recognition device; iris scanner; presence sensor; smart phone; smart watch; temperature sensor

B. Procedure

To understand how factors influence user privacy perception, we used the Generalized Mixed Effect Model (GLMM) with a random intercept per participant to construct our models. GLMM is a useful statistical approach for repeated measurement on the same person and very flexible when it comes to studying the interactions of different factors and their dependencies [14]. Using this approach allows us to find the best interactions between various factors, thus finding the best model in our research. GLMM also uses the Bayesian Information Criterion (BIC) for model selection. BIC, derived from Gideon Schwarz [15], is a widely used statistical approach that balances the number of parameters and data points against the maximum likelihood function and measures the efficiency of parameterized models when predicting the output. A lower BIC always indicates a better model where Δ BIC has to be above five between models to be considered a significant improvement. We used R programming with the Lme4 package to construct our models. We coded our responses in a binary format as follows: (Strongly Agree, Agree, Neutral $\rightarrow$ 1, Disagree, Strongly disagree $\rightarrow$ 0, Very Comfortable, Comfortable, Neutral $\rightarrow$ 1, Very Uncomfortable, Uncomfortable $\rightarrow$ 0). In contrast to the approach taken in the original study, we decided to unify our code scale across all our GLMM models. To clarify, our statistical models produced positive estimates for agreeing or being comfortable with data collection and negative estimates for disagreeing or being uncomfortable with such collections. This was not the case in the Allow/Deny model presented in the original study.

C. Qualitative analysis of responses

A qualitative examination was conducted on participants' responses to the open-ended questions in the study. Similar to the work that was conducted in the original study by Naeini et al. [6], participants' responses were coded with respect to five topics. These topics are as follows: the factors that the study's scenarios are based on, the aspects that contribute

TABLE III
DESCRIPTION OF THE INTER-ANNOTATOR AGREEMENT (IAA) BETWEEN THE TWO ANNOTATORS.

Categories	IAA
Factors	0.82
Whitelist	0.83
Blacklist	0.84
Control	0.83
Information	0.87
Risks	0.82

to enhancing the comfort level of participants, the aspects that impact the feeling of discomfort among participants, the information that participants would like to gain regarding data collection, and the ability to have control over their data, such as being able to delete the data or opt out at any time. As a result, a codebook was generated from all the answers that were given by participants in the study. Two annotators used the same codebook independently. The annotation process was performed on the Tagtog framework [16], [17]. Upon completing the annotation process, the Inter-Annotator Agreement (IAA) between the two annotators was provided by Tagtog as can be seen in table III. In terms of the categories' tags, the two annotators had an IAA that ranged from 0.57 to 1.0. Thus, we randomly selected one of the two annotator's coded responses and provided its findings in table IV.

IV. RESULTS

A. The Effects of Data Collection on Users' Comfort

Similar to the original study, participants were asked after each of the given scenarios to reflect their comfort level regarding their data being collected in that particular scenario via a five-point Likert scale that ranges from "Very Comfortable" to "Very Uncomfortable". The goal was to get a sense of understanding of the impact that each factor has on users' comfort. Figure 1 provides a presentation of the impact of these factors within their different levels. In regard to the data type factor, the majority of users were very uncomfortable when either biometric or video data was the type of data that is being collected in the given scenarios. Smartphone was the device type that has the most negative impact on the comfort level of participants as 54% were very uncomfortable when the data was collected through their smartphone. In respect to the location factor, most of the participants were very uncomfortable when public restroom was the place where the data collection occurred. In addition, analyzing the retention time factor revealed that users were comfortable when their collected data was only retained until the purpose for which the data was collected was satisfied. We also asked the participants to express their comfort level if their data was collected in each of the given scenarios, but we told the participants that they would not be able to know how long their data would be kept for, how long the data would be used and whether their data would be shared with another entity or not. We tried to examine how excluding the shared and retention time factors would impact the comfort level of our participants. As a result, the only significant changes are biometrics and department store being the most impactful factors for data type and location, respectively. In other words, participants were very uncomfortable with biometrics and department store when they are missing the retention time and data sharing.

B. Factors Affecting The Comfort Level

To analyze the data that we have acquired, we built a statistical model for the comfort level. Similar to the ones that had been built in the original study, the model was built using the generalized linear mixed model. In this model, the significance threshold is ($P < 0.05$) which means that any level of a factor that has a p value less than 0.05 would be considered statically significant. Moreover, we have built our models based on the BIC standard since it was the chosen standard in the original study and we wanted to compare our results to theirs. The model with the lowest BIC is the best model for describing the dependent variable, whereas the model with the highest BIC has the smallest effect on the dependent variable. Table V depicts the regression results that we have for the comfort level model. In our model, user perceived benefit was the factor that contributes the most to expressing the comfortableness of our participants with data collection. In other words, participants are more likely to feel comfortable with their data being collected whenever they view the collection process as beneficial to them. In this model, a positive estimate signals a leaning toward being comfortable with data collection, whereas a negative estimate signals a leaning toward discomfort with the process of data collection. Analyzing the coefficient of the factors' levels in the interaction between the location and data type factors reveals that participants are likely to feel uncomfortable with the collection of their biometric data at a public restroom. On the other hand, analyzing the outcome of the retention factor explains that participants are likely to feel comfortable with the data collection if their data is being stored until the purpose of the data collection is satisfied. Analyzing the outcomes of our models reveal that user perceived benefit, data type, and location are the main factors that had the most effect on the comfort level of our participants.

C. Factors Affecting Users' Allow/Deny Decisions

We built a regression model for the allow/deny decisions. In this model, which is presented in table VI, a positive estimate signals the likeliness to allow data collection, whereas a negative estimate signals the likeliness to deny the data collection. Analyzing the model shows that user perceived benefit is the factor that has the most impact on the participants' decisions to allow or deny the collection of their data. The allow/deny decisions were also highly affected by the interactions between data type and user perceived benefit. For instance, the results of our model allow us to state that participants are likely to deny the collection of their video data if they view the collection process as not beneficial to them. Besides, the interaction between data type and user benefit factors indicates that users'

TABLE IV
CATEGORIES AND CODES USED FOR THE CODEBOOK. PERCENTAGE IN BRACKETS REFLECTS HOW IMPORTANT EACH TAG IS FOR THE PARTICIPANTS.

Categories	**Tags(Importance)**
Factors	purpose(5.50%), data(55.96%), retention(2.75%), sharing(5.50%), benefit(8.26%), location(10.09%), device(3.67%)
Whitelist	safety(34.29%), common_good(14.29%), public(11.43%), personal_benefit(20%), improve_services(8.57%), anonymous_data(11.43%)
Blacklist	commercial(9.23%), everything(12.31%), personal_information(32.31%), private_location(3.08%), unknown_entities(13.85%), identifiable_information(23.08%), location(6.15%)
Control	consent(28%), ownership(28%), opt_out(12%), deletion(16%), access(16%)
Information	data_security(13.85%), purpose(18.46%), sharing_details(20%), retention(6.15%), data_handling(23.08%), collector(18.46%)
Risks	misuse(25.61%), intransparency(48.78%), personal_privacy(13.41%), surveillance(1.22%), tracking(1.22%), data_security(9.76%)

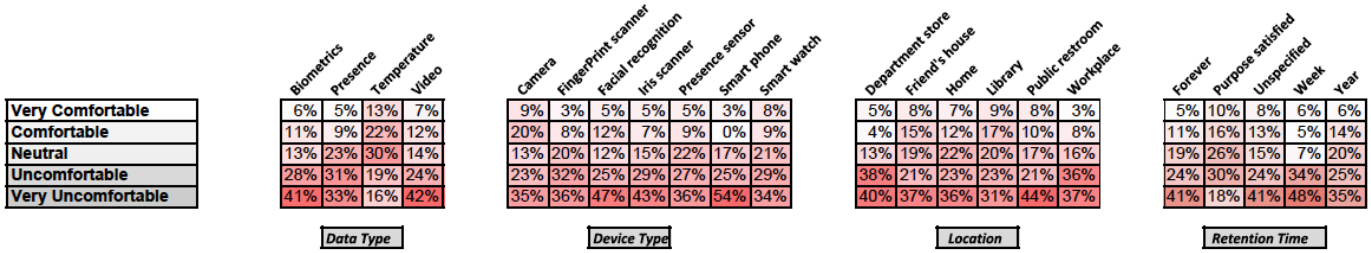

Fig. 1. Users' comfort with the data collection factors.

TABLE V
GLMM REGRESSION OUTCOME OF THE COMFORT LEVEL MODEL.

Factor	Estimate	Std. Err	Z-value	P-value	BIC
user perceived benefit					604.9
Baseline: beneficial					
not beneficial	−2.80	0.30	−9.26	0.00	
location:user perceived benefit					617.3
Baseline: workplace:beneficial					
public restroom:not beneficial	−3.71	1.23	−3.02	0.00	
location:data type					680.3
Baseline: temperature:library					
biometric:workplace	4.42	1.25	3.52	0.00	
biometric:public restroom	−6.57	2.17	−3.01	0.00	
video:public restroom	−8.7	2.07	−4.19	0.00	
presence:friend's house	−5.12	2.08	−2.45	0.01	
video:friend's house	−5.79	2.05	−2.81	0.00	
location:user benefit					687.5
Baseline: workplace:user					
collector:library	2.24	1.07	2.07	0.03	
both:library	2.68	1.09	2.46	0.01	
collector:public restroom	−6.73	1.89	−3.54	0.00	
collector:friend's house	3.44	1.77	−1.94	0.05	
both:friend's house	−4.65	1.80	−2.58	0.00	
collector:home	2.48	1.04	2.39	0.01	
data type					691.7
Baseline: temperature					
biometric	−2.31	0.37	−6.16	0.00	
presence	−1.77	0.38	−4.57	0.00	
video	−2.05	0.39	−5.27	0.00	
Retention					707.9
Baseline: unspecified					
forever	0.35	0.33	1.03	0.29	
until satisfied	1.34	0.39	3.40	0.00	
week	−1.08	0.44	−2.45	0.01	
year	0.81	0.33	2.43	0.01	
user benefit					708.5
Baseline: user					
both	0.42	0.27	1.51	0.12	
collector	−0.82	0.28	−2.93	0.00	

TABLE VI
GLMM REGRESSION OUTCOME OF THE ALLOW/DENY MODEL.

Factor	Estimate	Std. Err	Z-value	P-value	BIC
user perceived benefit					612.9
Baseline: beneficial					
not beneficial	−3.28	0.28	−11.56	0.00	
data type:user perceived benefit					626.4
Baseline: temperature:beneficial					
biometric:not beneficial	−1.57	0.90	−1.74	0.08	
presence:not beneficial	−0.03	0.88	−0.04	0.96	
video:not beneficial	−2.21	0.93	−2.37	0.01	
user perceived benefit:retention					654
Baseline: not beneficial:unspecified					
forever:beneficial	−1.50	0.72	−2.06	0.03	
until satisfied:beneficial	14.37	418.04	0.03	0.97	
week:beneficial	−0.45	0.92	−0.49	0.62	
year:beneficial	−0.36	0.79	−0.45	0.64	
user benefit					787.5
Baseline: user					
collector	−1.32	0.25	−5.18	0.00	
both	0.04	0.25	0.17	0.86	
data type:user benefit					789.9
Baseline: biometric:collector					
temperature:user	0.27	0.79	0.34	0.72	
presence:user	−1.56	0.59	−2.62	0.00	
temperature:both	2.46	1.09	2.25	0.02	
presence:both	−0.53	0.82	−0.64	0.51	
video:both	1.88	0.66	2.84	0.00	
data type:location					804.5
Baseline: workplace:temperature					
biometric:department store	−1.59	1.24	−1.27	0.20	
presence:department store	0.03	1.16	0.03	0.97	
biometric:library	−3.24	1.05	−3.07	0.00	
presence:library	−2.45	1.07	−2.27	0.02	
biometric:public restroom	−4.93	1.32	−3.71	0.00	
video:public restroom	−3.87	1.41	−2.72	0.00	
biometric:friend house	−2.63	1.33	−1.97	0.04	
presence:friend house	−1.39	1.48	−0.94	0.34	
video:friend house	−1.37	1.66	−0.82	0.41	

decisions to allow or deny the data collection are likely to depend on whether the data is sensitive to them, such as biometric data, and the benefits that users could gain from the collection process.

D. Users' Preferences Regarding The Frequency of Notifications

We wanted to study the effect of the notification system on users' data collection in different IoT scenarios using three different frequencies. Participants were requested to reveal

their preferences using a five point Likert scale from "Strongly agree " to "Strongly disagree". Their responses were then coded in a binary manner and cross-matched with various factors' levels utilized in previous GLMM models. The best models revealed that location, data type and user perceived benefit are the most influential factors on users willingness to get notified. Those results confirmed our codebook findings where participants stress on the need to be informed about the "Who", the "Where", and the "How" (Most dominant tags on the codebook). The analysis and model selection for the three notification systems are provided in the next sections.

E. Factors Affecting The Desire for Every-time Notifications

Using the GLMM regression model, we were also able to order various factors and their dependencies using the change in BIC to study their impact and contribution on user's willingness to get notified every time their data were collected. Table VII illustrates the results obtained from the top eight factors (ordered by their BIC size), where a positive coefficient indicates the likeliness of getting notified every time a data collection occurred. We found out that user perceived benefit was the most influential factor on users' desire to get notified while Location had the smallest impact. User perceived benefit factor aligned with our previous GLMM models ranking as well as with our codebook analysis, showing that participants perceive their personal benefit above all other aspects. Our statistical model also revealed that data type and location had the most significant dependencies. Analyzing various estimates among different factors affirms that participants want to be notified every time sensitive identifiable information is collected such as biometrics and video. The GLMM model also revealed that users are willing to be notified every time they are in a department store were sensors are continuously gathering data without any proper consent. Nonetheless, table VII implies that users are willing to deny notifications if the sharing details are mentioned and the collector is being transparent with their data.

TABLE VII
GLMM REGRESSION OUTCOME OF THE EVERY-TIME NOTIFICATION MODEL.

Factor	Estimate	Std. Err	Z-value	P-value	BIC
user perceived benefit					539.9
Baseline: beneficial					
not beneficial	0.87	0.29	2.91	0.00	
shared					544.7
Baseline: not shared					
shared	−0.89	0.45	−1.95	0.05	
data type					545.6
Baseline: temperature					
biometric	1.58	0.40	3.93	0.00	
presence	1.24	0.42	2.91	0.00	
video	1.07	0.41	2.58	0.00	
retention					545.7
Baseline: unspecified					
forever	0.37	0.40	0.93	0.35	
until satisfied	−0.34	0.45	−0.76	0.44	
week	2.09	0.59	3.53	0.00	
year	0.03	0.39	0.07	0.93	
user perceived benefit:happening within 2yrs					549.3
Baseline: beneficial:disagree					
not beneficial:agree	−1.23	0.77	−1.60	0.10	
user perceived benefit:happening today					551.5
Baseline: beneficial:disagree					
not beneficial:agree	0.81	0.74	1.08	0.27	
data type:user perceived benefit					555
Baseline: video:beneficial					
presence:not beneficial	−2.34	0.79	−2.94	0.00	
biometric:not beneficial	−1.05	0.73	−1.44	0.14	
temperature:not beneficial	−0.60	1.09	−0.55	0.58	
location					556.8
Baseline: library					
friend's house	0.30	0.38	0.78	0.43	
department store	1.82	0.51	3.55	0.00	
home	0.78	0.39	1.98	0.04	
public restroom	0.61	0.43	1.40	0.15	
workplace	1.06	0.43	2.48	0.01	

F. Factors Affecting The Desire for First-time Notifications

We also built a regression model that tackles users' inclination to get notified only for the first time when their data is being collected. In this model, which is shown in table VIII, a positive estimate reflects users' willingness to be notified for the first time of data collection, and a negative estimate implies the unwillingness of users to get such notification. Similar to the other statistical models that we built, factors were ordered in terms of their impact on users and their decision on whether to get notified about the occurrence of data collection when it happened for the first time. The outcomes of the model demonstrate that user perceived benefit is the most impactful factor that can express the inclination of users toward wanting to be notified for the first time of data collection. This observation is similar to the other observations that were made in previous models where user perceived benefit was found to be the most imperative factor. Additionally, the interaction between user perceived benefit and happening today factors allows us to state that users would like to be notified about the first occurrence of data collection. This leads us to hypothesize that users desire to be informed of the collection of their data even if the collection process doesn't s necessarily seem to be really happening in the near future. Moreover, our analysis shows that users are not willing to be informed only for the first time when the collection of their data involves benefiting the collector. This interesting observation may signal how users are suspicious of entities benefiting from the collection of their data and so they prefer to always be informed when the collection occurs as opposed to only getting notified once when the collection happens for the first time. This observation is in line with our qualitative analysis where participants echoed their fear and skepticism of the entities that collect their data.

G. Factors Affecting The Desire for Once-in-Awhile Notifications

A statistical model that analyzes users' desires to get notified every once in a while about their collected data was built. Table IX depicts the model and the outcomes of the factors that impacted the decisions of users. A positive estimate signals the tendencies of users to wishing to be notified every once in a while, whereas a negative estimate reflects the tendency of unwillingness to receive such notifications regarding data collection. Similar to the other models, user perceived benefit was also found to be the factor that best expresses users' desires and decisions. Nevertheless, location was found to be the factor that has the least effect on users' inclination towards wanting to be notified every once in a while about the occurrence of data collection.

TABLE VIII
GLMM REGRESSION OUTCOME OF THE FIRST-TIME NOTIFICATION MODEL.

Factor	Estimate	Std. Err	Z-value	P-value	BIC
user perceived benefit					648.7
Baseline: beneficial					
not beneficial	−0.86	0.24	−3.55	0.00	
user perceived benefit:happening today					654.6
Baseline: beneficial:not happening today					
not beneficial:not happening today	1.89	0.73	2.57	0.01	
user benefit					660.5
Baseline: user					
both	−0.74	0.30	−2.45	0.01	
collector	−0.67	0.29	−2.25	0.02	
data type					665.4
Baseline: temperature					
video	−1.00	0.37	−2.66	0.00	
biometric	−0.68	0.35	−1.94	0.04	
presence	−0.36	0.38	−0.96	0.33	
retention					674.8
Baseline: week					
unspecified	0.46	0.39	1.16	0.24	
year	0.37	0.36	1.03	0.30	
forever	0.68	0.36	1.86	0.06	
until satisfied	0.89	0.42	2.09	0.03	
location					676.5
Baseline: library					
workplace	0.11	0.34	0.32	0.74	
friend house	0.06	0.33	0.19	0.84	
department store	0.08	0.37	0.22	0.82	
home	0.67	0.34	1.97	0.04	
public restroom	−0.61	0.37	−1.64	0.09	

TABLE IX
GLMM REGRESSION OUTCOME OF THE ONCE-IN-AWHILE NOTIFICATION MODEL.

Factor	Estimate	Std. Err	Z-value	P-value	BIC
user perceived benefit					621.9
Baseline: beneficial					
not beneficial	−0.56	0.24	−2.29	0.02	
shared					626.5
Baseline: not shared					
shared	0.13	0.35	0.37	0.70	
user benefit					627.8
Baseline: user					
both	−0.73	0.31	−2.33	0.01	
collector	−0.59	0.30	−1.93	0.05	
user perceived benefit:happening within 2 years					629.3
Baseline: beneficial:not happening within 2 yrs					
not beneficial:happening within 2 yrs	−1.45	0.00	−1449.4	0.00	
user perceived benefit:happening today					634.5
Baseline: beneficial:not happening today					
not beneficial: happening today	0.03	0.66	0.05	0.95	
Retention					640.8
Baseline: unspecified					
week	0.06	0.40	0.15	0.87	
year	0.45	0.34	1.30	0.19	
forever	0.62	0.35	1.77	0.07	
until satisfied	−0.04	0.40	−0.10	0.91	
location					645.7
Baseline: workplace					
library	0.12	0.36	0.34	0.73	
friend's house	0.39	0.38	1.01	0.30	
department store	0.46	0.42	1.10	0.26	
home	0.75	0.38	1.94	0.05	
public restroom	−0.19	0.41	−0.46	0.64	

V. DISCUSSION

We present our observations after analyzing the results mentioned in Section IV. Like the original study, biometric had a more negative impact on our participants' comfort level than environmental data such as the collection of presence. However, unlike the original study, collecting biometric data is not the most dominant data type affecting participants' comfort level as the collection of video data had a slight more impact with 41% for the former and 42% for the latter. Our study conducted five years apart from the original research also revealed somewhat different results, especially toward smartphones (54% vs. 25%), and workplace (37% vs. 17%). We associate this outcome with the growing number of recurrent leaked private videos of users in the past few years, including the incident that affected numerous Google's users in late 2019 [18], [19]. In the original study, participants were very uncomfortable with iris scanner as a device type, whereas, in our research, smartphones were the most dominant device type. This outcome can be tied to how smartphones have become such an integral part of users' lives where they can be considered by users as their virtual friends or even extensions of themselves [20], [21]. In fact, the smartphone industry has grown exponentially over the past few years and incorporated various sensing technology such as fingerprint scanners, facial/iris recognition, and tracking, which made users' sensitive information exposed to third party partners [22].

Through the analysis of our GLMM models, we found user perceived benefit to be the factor that most expressed participants' comfort level and their allow or deny decisions. this finding was in line with an observation made in the original study where user perceived benefit was found to be greatly vital in users' decisions on data collection and their level of comfort regarding that collection. The importance of data type and location factors in expressing users' privacy preferences was also a shared finding between our study and the original one. Nonetheless, one difference in the results that we gathered compared to the original study was the effects of the shared factor. In our study, participants were more comfortable when they were told to whom the data would be shared with. This is in line with prior research recommendations and recent data protection frameworks such as General Data Protection Regulation (GDPR) that call for transparency in data collection practices [23], [24]. Traces of calling for honesty and trust was also found in our codebook analysis where participants stress the need of clarity between users and collectors. Unlike the original study that found the greatest factors in expressing the comfort level and allow/deny decisions of participants to be different, our results concluded that user perceived benefit is dominant in our GLMM models. This finding is similar to prior research that highlighted the impact of users' perception on their privacy preferences and decisions [1], [25], [26].

Concerning the notification system, we noticed somewhat similar findings between both studies. User perceived benefit was a common factor that impacted user's willingness to get notified as well as data type and location. In our study the every-time notification model contained more dependencies compared to other notification models. We hypothesize that participants would like to be notified whenever their sensitive information or location is being collected to have some sense of awareness, giving them some degree of freedom and choice.

VI. LIMITATIONS AND FUTURE WORK

Our study contributes to the growing body of work that investigates user privacy preferences [1], [24], [25] by conducting a replication study of a larger vignette [6] Although our survey is limited to fewer participants, it echoes the findings of the original study that users' perceived benefits is an essential and motivating factor. Despite the limited participants, we observe an interesting deviation from the original survey that demands future work. Our work identified

the workplace has now morphed into a sensitive location where users are uncomfortable sharing their private information. Similar to the original study, our study may suffer from the privacy paradox phenomenon as a result of our participants expressing their privacy preferences through responding to the hypothetical scenarios that were given to them [6]. The action of a participant could differ from their response in the study if they were actually experiencing a situation similar to one of the given hypothetical scenarios. In addition, while we did not intend for this study's gender population to be skewed towards one gender, our findings might be influenced by the gender population being mainly of the male gender. Nonetheless, these findings lay the path for further investigation on the impact of gender differences on the privacy preferences of users in an IoT environment, which we will reserve for a future study.

VII. Conclusion

To understand users' privacy preferences, we conducted a replication study that asked participants to imagine themselves in situations where IoT devices collect their data. Our participants' responses helped us determine the factors that could affect participants' comfort level, their decisions to either allow or deny the process of data collection, and how frequently they are willing to get notified in such scenarios.

We found that users' perceived benefit is an important factor that may motivate them to feel comfortable and allow data collection to be favorable. Dissimilar to the original study, we found out that participants' consider the workplace as a sensitive location where they feel uncomfortable sharing their private information. We hope that our findings combined with the original study's would enrich the literature regarding privacy preferences and further enlighten manufacturers on how to collect data when designing their products.

Acknowledgement

This material is based upon work supported in whole or in part with funding from the Office of Naval Research (ONR) contract #N00014-20-1-2798. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the ONR and/or any agency or entity of the United States Government.

References

[1] D. Kim, K. Park, Y. Park, and J.-H. Ahn, "Willingness to provide personal information: Perspective of privacy calculus in iot services," *Computers in Human Behavior*, vol. 92, pp. 273–281, 2019.

[2] J. Deogirikar and A. Vidhate, "Security attacks in iot: A survey," in *2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*. IEEE, 2017, pp. 32–37.

[3] L. Atzori, A. Iera, and G. Morabito, "The internet of things: A survey," *Computer networks*, vol. 54, no. 15, pp. 2787–2805, 2010.

[4] R. Chow, "The last mile for iot privacy," *IEEE Security & Privacy*, vol. 15, no. 6, pp. 73–76, 2017.

[5] L. Sagnières, "Nissenbaum, h., privacy in context: Technology, policy, and the integrity of social life," 2013.

[6] P. E. Naeini, S. Bhagavatula, H. Habib, M. Degeling, L. Bauer, L. F. Cranor, and N. Sadeh, "Privacy expectations and preferences in an iot world," in *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*, 2017, pp. 399–412.

[7] B. Könings, S. Thoma, F. Schaub, and M. Weber, "Pripref broadcaster: Enabling users to broadcast privacy preferences in their physical proximity," in *Proceedings of the 13th International Conference on Mobile and Ubiquitous Multimedia*, 2014, pp. 133–142.

[8] L. Tsai, P. Wijesekera, J. Reardon, I. Reyes, S. Egelman, D. Wagner, N. Good, and J.-W. Chen, "Turtle guard: Helping android users apply contextual privacy preferences," in *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*, 2017, pp. 145–162.

[9] B. P. Knijnenburg, "Information disclosure profiles for segmentation and recommendation," in *SOUPS2014 Workshop on Privacy Personas and Segmentation*, 2014.

[10] B. Liu, M. S. Andersen, F. Schaub, H. Almuhimedi, S. A. Zhang, N. Sadeh, Y. Agarwal, and A. Acquisti, "Follow my recommendations: A personalized privacy assistant for mobile app permissions," in *Twelfth Symposium on Usable Privacy and Security ({SOUPS} 2016)*, 2016, pp. 27–41.

[11] J. Lin, B. Liu, N. Sadeh, and J. I. Hong, "Modeling users' mobile app privacy preferences: Restoring usability in a sea of permission settings," in *10th Symposium On Usable Privacy and Security (SOUPS 2014)*, 2014, pp. 199–212.

[12] H. Lee and A. Kobsa, "Privacy preference modeling and prediction in a simulated campuswide iot environment," in *2017 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 2017, pp. 276–285.

[13] [Online]. Available: https://www.notion.so/

[14] J. J. Faraway, *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. CRC press, 2016.

[15] A. A. Neath and J. E. Cavanaugh, "The bayesian information criterion: background, derivation, and applications," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 4, no. 2, pp. 199–203, 2012.

[16] J. M. Cejuela and B. Rost, "tagtog: collaborative interactive semi-supervised learning and annotation web-based framework," 2011.

[17] J. M. Cejuela, P. McQuilton, L. Ponting, S. J. Marygold, R. Stefancsik, G. H. Millburn, B. Rost, F. Consortium *et al.*, "tagtog: interactive and text-mining-assisted annotation of gene mentions in plos full-text articles," *Database*, vol. 2014, 2014.

[18] "Google admits it sent private videos in Google Photos to strangers - The Verge," https://www.theverge.com/2020/2/4/21122044/google-photos-privacy-breach-takeout-data-video-strangers.

[19] "Google Guilty Of 'Big Screw Up' That May Have Leaked Your Videos To A Random Stranger - Forbes," https://www.forbes.com/sites/thomasbrewster/2020/02/04/google-photos-makes-big-screw-up-and-mayve-leaked-your-videos-to-a-random-stranger/?sh=79c961db5486.

[20] C. Fullwood, S. Quinn, L. K. Kaye, and C. Redding, "My virtual friend: A qualitative analysis of the attitudes and experiences of smartphone users: Implications for smartphone attachment," *Computers in Human Behavior*, vol. 75, pp. 347–355, 2017.

[21] A. Carolus, J. F. Binder, R. Muench, C. Schmidt, F. Schneider, and S. L. Buglass, "Smartphones as digital companions: Characterizing the relationship between users and their phones," *New Media & Society*, vol. 21, no. 4, pp. 914–938, 2019.

[22] Y. Wang, Y. Chen, F. Ye, H. Liu, and J. Yang, "Implications of smartphone user privacy leakage from the advertiser's perspective," *Pervasive and Mobile Computing*, vol. 53, pp. 13–32, 2019.

[23] "GDPR Archives - GDPR.eu," https://gdpr.eu/tag/gdpr/.

[24] M. Tabassum, T. Kosinski, and H. R. Lipford, "" i don't own the data": End user perceptions of smart home device data practices and risks," in *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*, 2019.

[25] I. Psychoula, D. Singh, L. Chen, F. Chen, A. Holzinger, and H. Ning, "Users' privacy concerns in iot based applications," in *2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*. IEEE, 2018, pp. 1887–1894.

[26] J. Lau, B. Zimmerman, and F. Schaub, "Alexa, are you listening? privacy perceptions, concerns and privacy-seeking behaviors with smart speakers," *Proceedings of the ACM on Human-Computer Interaction*, vol. 2, no. CSCW, pp. 1–31, 2018.