Privacy Expectations and Preferences in an IoT World

Pardis Emami Naeini, Sruti Bhagavatula, Hana Habib, Martin Degeling, Lujo Bauer, Lorrie Cranor, and Norman Sadeh, *Carnegie Mellon University*

https://www.usenix.org/conference/soups2017/technical-sessions/presentation/naeini

This paper is included in the Proceedings of the
Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017).

July 12–14, 2017 • Santa Clara, CA, USA

ISBN 978-1-931971-39-3

Privacy Expectations and Preferences in an IoT World

Pardis Emami-Naeini, Sruti Bhagavatula, Hana Habib, Martin Degeling, Lujo Bauer,
Lorrie Faith Cranor, Norman Sadeh
Carnegie Mellon University
Pittsburgh, PA, USA

ABSTRACT

With the rapid deployment of Internet of Things (IoT) technologies and the variety of ways in which IoT-connected sensors collect and use personal data, there is a need for transparency, control, and new tools to ensure that individual privacy requirements are met. To develop these tools, it is important to better understand how people feel about the privacy implications of IoT and the situations in which they prefer to be notified about data collection. We report on a 1,007-participant vignette study focusing on privacy expectations and preferences as they pertain to a set of 380 IoT data collection and use scenarios. Participants were presented with 14 scenarios that varied across eight categorical factors, including the type of data collected (e.g. location, biometrics, temperature), how the data is used (e.g., whether it is shared, and for what purpose), and other attributes such as the data retention period. Our findings show that privacy preferences are diverse and context dependent; participants were more comfortable with data being collected in public settings rather than in private places, and are more likely to consent to data being collected for uses they find beneficial. They are less comfortable with the collection of biometrics (e.g. fingerprints) than environmental data (e.g. room temperature, physical presence). We also find that participants are more likely to want to be notified about data practices that they are uncomfortable with. Finally, our study suggests that after observing individual decisions in just three data-collection scenarios, it is possible to predict their preferences for the remaining scenarios, with our model achieving an average accuracy of up to 86%.

1. INTRODUCTION

The Internet of Things (IoT), composed of network-connected physical objects, is growing rapidly. The devices that make up the IoT vary greatly in their form and purpose, from sensors that people voluntarily carry on their wrists, to network-connected thermostats, to street lights that count the number of people who pass by. While these devices bring about new services, increase convenience, and improve efficiency, they also bring privacy and security risks.

To fully realize the potential of IoT, individuals need to be sufficiently knowledgeable and aware to make informed decisions. Thus, IoT devices need to inform their users about their data collection practices and offer privacy choices that respect individual privacy preferences. Gaining traction on this problem requires nuanced understanding of societal norms and context, as well as individual needs [31, 35]. For example, most people tacitly accept being recorded on cameras and CCTV outdoors in public spaces, but express disdain for installing video surveillance systems inside the walls of their homes. As more complex IoT scenarios become possible, many other factors may play a role in determining individuals' privacy preferences. While some may feel comfortable with their location being tracked for the purpose of traffic prediction, they may consent to tracking only their work commute. Others may consent only if they are assured that their location data is retained and used in an anonymized form.

We conducted a large-scale online vignette study to identify the contribution of different factors (such as the type of data, retention time, purpose of data collection, and location of data collection) in promoting or inhibiting individuals' self-professed comfort levels. We also studied the factors that trigger a desire for notifications about data collection. Our research identified which aspects of data collection or use by various IoT devices are most likely to cause discomfort, how realistic participants think these scenarios are, and which aspects they would like to be made aware of.

The results of our study informs the design of more transparent IoT-connected systems—we envision our results can be used to improve privacy notices for IoT devices, and develop more advanced personal privacy assistants [25].

This paper makes two main contributions. First, we show that individuals' comfort levels in a variety of IoT data collection scenarios are related to specific aspects of that data collection. Many of our findings are consistent with observations made in prior work, but our quantitative methodology and the scale of our experiment allows us to understand the effect of individual factors and their relative importance more precisely. Second, leveraging our qualitative and quantitative results, we advance explanations for many of the differences among these factors. We show that whether or not participants think the use of their data is beneficial to them has a profound influence on their comfort level. We also find that participants' desire for notification is closely related to whether or not they feel comfortable with data collection in a particular scenario.

The paper is organized as follows. First, we discuss related work. Then we describe the design of our vignette study, and discuss our quantitative and qualitative analysis of our survey data. Next, we present the results of our prediction model, and draw conclusions from the analysis. Finally, we discuss study limitations and possible approaches to mitigate some of the concerns highlighted by our study.

Symposium on Usable Privacy and Security (SOUPS) 2017, July 12–14, 2017, Santa Clara, California.

2. RELATED WORK

Our research builds on prior work aimed at understanding individuals' IoT-related privacy concerns, and potential solutions for mitigating them [6, 8]. Additionally, prior research has studied various factors that can impact privacy preferences, the results of which were used to inform the design of our study. Recent work has also developed models to predict individuals' privacy preferences, so that data collection can be personalized to suit people's privacy preferences. Our work aims to address privacy concerns in a variety of IoT scenarios where sensing is pervasive. Our work underscores the relative importance of different privacy concerns to individuals. These findings inform the understanding of privacy preferences as they relate to IoT data collection.

2.1 IoT Privacy Challenges

New methods of data collection in the IoT have led to new privacy challenges. Some of these challenges include obtaining consent for data collection, allowing users to control, customize, and choose the data they share, and ensuring the use of collected data is limited to the stated purpose [33]. These challenges are made more difficult by the increased potential for misuse of personal information in the IoT domain. This stems from the pervasive tracking of habits, behaviors, and locations over a long period of time. There are new risks to personal safety introduced by IoT systems [6, 9]. Addo et al. demonstrated that trustworthiness of an IoT application is impacted by the implemented privacy and security practices [2]. To be accepted by consumers, IoT-connected device developers must consider the privacy and security implications of their products.

2.2 Privacy Interfaces for IoT Systems

There have been several proposals to help address privacy concerns related to data collection in the IoT domain. Mehrotra et al. presented two systems that could help highlight privacy challenges associated with IoT sensing and allow for testing of various privacy-enhancing solutions [30]. Lederer et al. identified five "pitfalls" in designing systems, particularly in ubiquitous computing environments, which lead to negative implications for individual privacy [18]. To address some of these pitfalls, Egelman et al. used crowdsourcing techniques to study different designs of privacy icons for a camera, with the aim of helping individuals make an informed decision about their privacy. Though many of their icons were successful in conveying what data was being collected, many participants demonstrated low comprehension. These findings underscored the difficulty of successfully informing individuals about what is going on around them in an IoT setting [12]. Recognizing the privacy risk caused by involuntary disclosure of information in IoT environments, Ukil et al. proposed a privacy management scheme that estimates a domain-specific measure of risk due to privacy disclosure in smart energy applications [38].

According to Bhaskar et al., a major limitation of prior work studying privacy in IoT environments is that studies typically focus on a single environment in which IoT sensing is occurring [6]. Thus, many of the proposed solutions do not generalize to other IoT contexts. Our work attempts to address this shortcoming by identifying privacy concerns in multiple heterogeneous scenarios which employ different types of data collection. This way, our methodology can determine which factors have the greatest impact on measures of individuals' comfort with data collection. The results can inform the design of privacy-enabling solutions appropriate to the variety of contexts we have studied. Furthermore, our study aims to expand beyond prior work in this area by identifying privacy concerns individuals have in data collection scenarios which are not obviously aligned with specific privacy risks.

2.3 Factors Impacting Privacy Preferences

Prior studies outside of the IoT context have examined different factors that can impact individuals' willingness to share information, based on measures of comfort with data collection. Bilogrevic et al. found that the comfort levels associated with sharing data are highly dependent on the specific type of data and the sharing context (e.g. search engines, social networks, or online shopping sites) [7]. Leon et al. tested whether data retention, access to collected information, and the scope of use affected willingness to share data for online behavioral advertising purposes. Individuals were more willing to share certain types of data if it had a retention period of one day, but for periods longer than one week, individuals were less likely to be willing to share [22].

Other work has focused on privacy preferences related to mobile devices and applications. Lin et al. evaluated individuals' perceptions of requests to access privacy-sensitive resources (e.g. sensors) on mobile devices. They found that both individual expectations of what an app does and the purpose for which an app requests access to sensitive resources impacts their privacy decisions [23]. In order to better understand people's attitudes toward sharing their location in mobile applications, Sadeh et al. built a system that enabled mobile device users to select and limit with whom they want to share their location. They concluded that increasing people's awareness has a critical role in helping them define more precise policies for protecting their privacy [36]. Tsai et al. studied the impact of giving feedback to mobile device users. Their study informed participants about who their data is being shared with, and when the data was shared. The goal was also to help people manage their privacy on a location sharing application. They reported that when people get adequate feedback, they are more willing to share data. They were also more comfortable with sharing their location [37].

Other studies more closely aligned with our work have evaluated several factors that may impact privacy concerns related to IoT data collection. Lederer et al. studied the relative importance of two factors; the entity collecting data, and the situation in which it is being collected, for determining users' privacy preferences in ubiquitous computing settings. Their results indicate that individuals base their privacy decisions on who is collecting their data, rather than the context in which it is being collected [19]. Lee and Kobsa tested five factors related to the context of data collection in two separate studies and found that individuals generally thought that monitoring in personal spaces was unacceptable, along with monitoring by an unknown entity or the government. Their results also indicate that photo and video monitoring may cause some privacy concern regardless of context [20, 21]. Other small, qualitative studies have focused on individuals' privacy preferences related to wearable sensors. These studies revealed that people demand ownership of the data they produce, and that privacy concerns vary depending on factors including retention time and the perceived value of the data collected [4, 17].

Our work leverages prior work to identify several factors that may impact individuals' privacy concerns and preferences in IoT settings. While data retention was found to be a significant factor in an online context [22], we aim to determine whether this remains true for IoT data collection. Additionally, the impact of the location of the data collection, type of data being collected, and purpose for collection have already been studied in prior work considering IoT contexts [20, 21]. We aim to expand on these findings by evaluating these factors in a larger scale study, and in combination with additional factors capturing more contextual nuances that are specific to IoT environments.

2.4 Predicting Privacy Preferences

Prior work has shown that privacy preferences can be inferred by segmenting collections of individuals based on profiles. These profiles represent clusters of different individuals and their privacy decisions. In the mobile app privacy domain, Lin et al. and Liu et al. demonstrated that a small number of profiles may be capable of predicting individuals' decisions to allow, deny, or be prompted for app permissions with a high level of accuracy [24, 26]. In IoT data collection scenarios, Lee and Kobsa were able to identify four clusters of participants with distinctive privacy preferences. These clusters were used to predict their study participants' decision to allow or deny monitoring in a particular IoT context with 77% accuracy [21]. In our work, we incorporate additional factors into a larger scale study, using similar techniques to make predictions with the goal of achieving improved prediction accuracy relative to prior work.

3. METHODOLOGY

We conducted a within-subjects survey with 1,014 Amazon Mechanical Turk [1] workers in order to understand individuals' privacy preferences. We exposed each participant to 14 different vignettes presenting an IoT data collection scenario. Vignettes are "short stories about hypothetical characters in specified circumstances, to whose situation the interviewee is invited to respond," [13] and have been used in prior work studying varying privacy contexts [28, 29].

Between vignettes, we varied eight factors that we hypothesized could influence individuals' privacy preferences:

- the type of data collected (data_type),
- the location where the data is collected (location),
- who benefits from the data collection (user_benefit),
- the device that collects the data (device_type),
- the purpose of data collection (purpose),
- the retention time (retention),
- whether the data is shared (shared), and
- whether additional information could be inferred from the collected data (inferred).

Several of these factors have already been shown in prior work to be important to individuals, when presented individually or in combination [4, 17, 19, 20, 21, 22]. Our design allowed these factors to be studied simultaneously, capturing more contextual nuances. In our vignettes, some factors could take on one of many possible levels. For reference, table 1 describes the factors and their corresponding levels.

After accepting the MTurk HIT, each study participant was directed to a survey where they were shown 14 different vignettes.

Each vignette introduced the factors being tested in the same order. In each scenario, vignettes began with the location of the data collection and ended with the retention period. The following is an example of a scenario presented to participants:

> You are at **work** and your **smart watch** is keeping track of your **specific position in the building**. Your position is shared with the **device manufacturer** to **determine possible escape routes in the case of an emergency or a hazard**. This data will be kept by the manufacturer **until you leave for the day**.

All factorial combinations of the different levels of each factor produced 126,720 possible scenarios, many of which contained combinations of factors which did not make sense (e.g. a presence sensor taking iris scans for emergency purposes). These scenarios were removed from the set of scenarios shown to participants. From the remaining set, we selected 380 scenarios that could feasibly occur, and ensured that this subset contained scenarios in which each level of each factor was represented. 14 vignettes drawn from these 380 scenarios so as to not overburden them. Randomly selecting subsets of 14 scenarios could have caused interaction effects due to a lack of diversity in each factor (e.g., presenting only one retention time on otherwise diverse scenarios) [3]. To minimize such interaction effects, we carefully selected subsets of vignettes so that every level of every factor was present at least once per subset, with the exception of the factors device_type, purpose, and inferred, which were dependent on other factors such as location, device_type, and user_benefit. In doing so, we divided the list of scenarios into 39 subsets with 14 scenarios each, and presented each participant with vignettes corresponding to one of these 39 subsets. The subsets were not mutually exclusive.

For each scenario, participants were asked how comfortable they were with data collection in that scenario and whether they found the use of data in the scenario to be beneficial (user_perceived_benefit). This factor is different from user_benefit, which refers to whether the data collection benefits the participant or the collector and is part of the scenario design; user_perceived_benefit refers to the participant's perception of whether the scenario would be beneficial to them. This question was only asked about scenarios in which a purpose was given; we coded this factor as 'N/A' for scenarios without a purpose. We also asked participants whether they would allow the data collection described in the scenario, and how often they would like to be informed about the data collection. Further questions asked how realistic a scenario was ("I think scenarios like this happen today," "... will happen within 2 years," and "... will happen within 10 years") and coded the answers to these three questions as happening_today, within_two_years, and within_ten_years, respectively. These three questions were answered on a five-point Likert scale from "Strongly Disagree" to "Strongly Agree" and were binned into binary categories based on agreement—0 (strongly disagree, disagree) and 1 (strongly agree, agree, neither agree nor disagree). Finally, we asked participants general demographic questions, followed by ten questions from the Internet Users' Information Privacy Concerns (IUIPC) scale to gauge their level of privacy concern. The IUIPC scale questions focus on concerns about control, awareness, and collection [27]. The complete set of questions asked in our survey is included in the Appendix.

3.1 Factors Impacting Preferences

We were interested in learning what factors of data collection contributed most significantly to individuals' comfort and preferences. Thus, we asked questions about how comfortable they were with the given scenario. We also asked if they would allow a specific data collection or not, and how often they would want to be notified about it. Participants' responses to these questions enabled us to build models that predict the concerns and preferences of the general population, based on our sample. We constructed five statistical models, capturing five dependent variables: comfort level, allow or deny decisions for the data collection, desire to be notified of data collection every time, desire to be notified once in a while, and desire to be notified only the first time. In addition to the eight factors in Table 1, we included the factors user_perceived_benefit, happening_today, within_two_years, within_ten_years, gender, age, income, and education, as well as the three IUIPC scale factors IUIPC-control, IUIPC-awareness, and IUIPC-collection.

[1] Amazon's Mechanical Turk https://www.mturk.com

Factor	Levels	Description
location	department store; library; workplace; friend's house; home; public restroom	location where the data is collected
data_type	presence; video; specific position; biometric data (e.g., fingerprint, iris, face recognition)	type of data collected
device_type	smart watch; smart phone; camera; presence sensor; temperature sensor; fingerprint scanner; facial recognition system; iris scanner	device that is collecting the data; some devices like smart phones can collect multiple data types
user_benefit	user (e.g., get help in emergency situations); data collector (e.g., downsize staff)	who benefits from the data collection and use
purpose	a specific purpose is mentioned; it is mentioned that participants are not told what the purpose is	purpose of data collection depends on the location, the data and who is benefiting
retention	forever; until the purpose is satisfied; unspecified; week; year	the duration for which data will be kept
shared	shared (e.g., with law enforcement); no sharing is mentioned	whether the data is shared or not
inferred	inferred (e.g., movement patterns); inferred data is not mentioned	Additional information can be inferred and users can be deanonymized

Table 1: Factors varied between vignette scenarios, levels of the factors presented in scenarios, and description of each factor.

We represented income as a quantitative variable based on categories of income ranges, excluding two outliers—participants who reported earning more than $200,000. We mapped all Likert scale responses to binary categories of 0 and 1, where 1 implies a positive preference, and 0 implies a negative preference. All of the quantitative variables (income, age, IUIPC-control, IUIPC-awareness, IUIPC-collection) were normalized before analysis to be on the same scale with a mean of 0 and standard deviation of 1.

We did not include two of the eight privacy factors, device_type and purpose. The device that is collecting the data was mentioned in the vignettes to make them more realistic, but was not considered in the statistical analysis because the device was uniquely determined by the type of data that was collected. The type of data that was collected was considered in the statistical analysis, resulting in a dependency between the two factors. Dependencies of this type between factor levels can lead to inaccurate statistical inferences. To improve the accuracy of our results, we excluded them from our statistical analysis. For the same reason, we removed purpose as it was not linearly independent from multiple other factors, such as location and user_benefit. Treating it as an independent factor would have resulted in scenarios that did not make sense contextually. For instance, using purpose as an independent factor would have included scenarios which involved collecting fingerprints to downsize staff. To eliminate these nonsensical scenarios from our study, we chose to remove purpose from the analysis, instead of the other factors on which it depended.

After removing these two factors, we found one of the subsets of scenarios contained two scenarios that differed only in these two factors. Therefore, for participants who received this subset, we removed the first of the two scenarios' answers and analyzed the remaining 13 scenarios.

Our models were constructed using generalized linear mixed model (GLMM) regression with a random intercept per participant. GLMM is particularly useful for modeling repeated measures experiments, such as ours, in which participants are presented with multiple parallel scenarios [5].

We performed model selection to find the best combination of factors by using a search algorithm with a backwards elimination approach. For each of our dependent variables, we found the model that best fit the data according to the Bayesian Information Criterion (BIC). We eliminated the variables with the largest p-value in each step of the model selection and continued the elimination until the BIC reached the global minimum [15]. The model with the lowest BIC best explains the dependent variable.

We present the regression tables for our best models in the Results section. We used a significance threshold of 0.05 to determine whether or not a factor was significant. Effects and the effect size of a factor level can be interpreted as proportional to the magnitude of the estimate co-efficient. We also defined a baseline for each factor. The regression tables and co-efficients of levels in the model were computed against the corresponding factors' baseline. Some of the baselines were selected based on specific concerns highlighted by our qualitative data, such as data_type (baseline = specific position) and location (baseline = friend's house). The baselines for other factors were selected based on their alphabetical ordering.

3.2 Predicting Preferences

Using the results from the model selection for each dependent variable, we further examined their predictive ability for individuals' preferences. Specifically, in our analysis we focus on predicting:

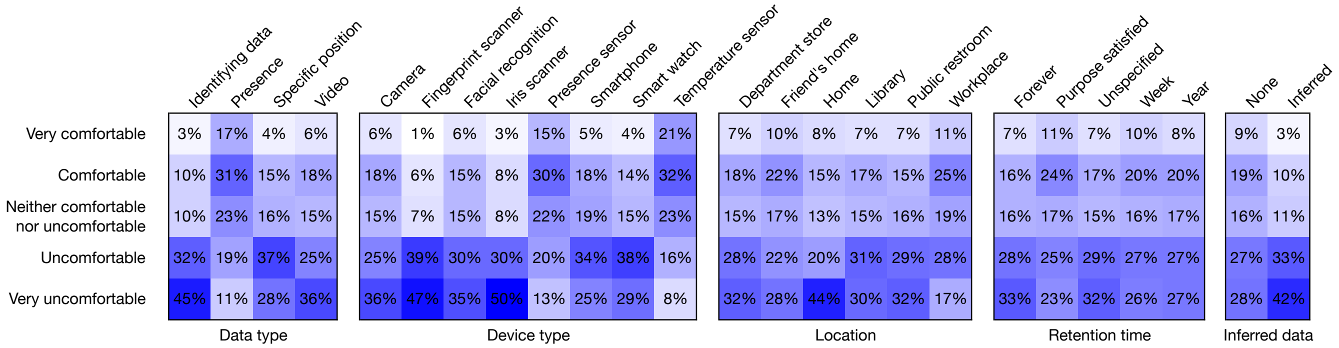
- an individual's comfort with a specific data collection scenario; and
- an individual's decision to allow or deny a specific data collection instance.

We believe that the ability to predict individuals' preferences or decisions is useful, since we can imagine deployment scenarios where a system needs to predict an individual's comfort or decision to allow or deny data collection. In these cases, the system would have more data accumulated over time specific to an individual using the system, and so would likely perform better than the classifiers in our experiments.

3.2.1 Features

For each of the two prediction tasks mentioned above, we used the main factors and interactions from the results of our model selection to predict the two outcomes; comfort level, and the decision to allow or deny.

Continuous features were encoded as-is in the feature vector, while categorical features were encoded as one-hot vectors for each category in the domain of that feature. This means, that each categorical variable was encoded as a vector of binary features where each feature corresponded to the binary value of one of the categories in the original categorical variable. In a one-hot vector, only one value in the whole vector will be 1 at any given time. This is a common way of encoding multi-class categorical features for machine learning tasks. For each categorical variable, the overall feature vector was



	Data type				Device type								Location						Retention time					Inferred data	
	Identifying data	Presence	Specific position	Video	Camera	Fingerprint scanner	Facial recognition	Iris scanner	Presence sensor	Smartphone	Smart watch	Temperature sensor	Department store	Friend's home	Home	Library	Public restroom	Workplace	Forever	Purpose satisfied	Unspecified	Week	Year	None	Inferred
Very comfortable	3%	17%	4%	6%	6%	1%	6%	3%	15%	5%	4%	21%	7%	10%	8%	7%	7%	11%	7%	11%	7%	10%	8%	9%	3%
Comfortable	10%	31%	15%	18%	18%	6%	15%	8%	30%	18%	14%	32%	18%	22%	15%	17%	15%	25%	16%	24%	17%	20%	20%	19%	10%
Neither comfortable nor uncomfortable	10%	23%	16%	15%	15%	7%	15%	8%	22%	19%	15%	23%	15%	17%	13%	15%	16%	19%	16%	17%	15%	16%	17%	16%	11%
Uncomfortable	32%	19%	37%	25%	25%	39%	30%	30%	20%	34%	38%	16%	28%	22%	20%	31%	29%	28%	28%	25%	29%	27%	27%	27%	33%
Very uncomfortable	45%	11%	28%	36%	36%	47%	35%	60%	13%	25%	29%	8%	32%	28%	44%	30%	32%	17%	33%	23%	32%	26%	27%	28%	42%

Figure 1: Summary statistics showing the relation between various factors and participants' comfort level. For example 45% of participants were very uncomfortable when the type of data being collected was biometric. Cells with larger numbers are darker in background color.

Gender		Age		Education		Income		IUIPC Score	
Male	49.2% (49.2%)	Range	18-78	No high school	0.8% (10.9%)	< $15k	16.4% (11.6%)	*Control Factor*	
Female	50.1% (50.8%)	Mean (SD)	36.1 (10.9)	High school	30.8% (28.8%)	$15k-$34k	33.8% (20.5%)	Range	1.33-7
No answer	0.7% (0.0%)	US average	37.9	Associates	9.7% (10%)	$35k-$74k	36.1% (29.4%)	Mean [SD]	5.95 [0.90]
				Bachelors	49.0% (48.7%)	$75k-$149k	9.3% (26.2%)	*Awareness Factor*	
				Professional	8.5% (1.5%)	$150k-$199k	0.9% (6.2%)	Range	1-7
				No answer	1.0% (0.0%)	> $200k	0.2% (6.1%)	Mean [SD]	6.44 [0.82]
						No answer	3.2% (0.0%)	*Collection Factor*	
								Range	1-7
								Mean [SD]	5.79 [1.11]

Table 2: Demographic breakdown of our participants. In the Gender, Education, and Income columns, the numbers in parentheses show the US average, according to census data from 2015.

increased in size by the size of each one-hot vector. For interactions between whole factors, we computed the product of each combination of the values in the one-hot vector and appended this vector of interaction products to the feature vector.

3.2.2 *Classifiers*

We experimented with various binary classifiers for the allow/deny prediction, and both binary and continuous classifiers for the comfort prediction. For binary classifiers where the outcome is binary, we used logistic regression, support vector machines (SVM), k-Nearest Neighbor, AdaBoost (with various weak base classifiers), and simple neural networks in the form of three-layer multi-layer perceptrons (MLP) [32]. For predicting comfort, we also experimented with a continuous version of the comfort level on a scale from 1 to 5, normalized to be between 0 and 1, for which we used linear regression for prediction.

We found the AdaBoost classifier with a logistic regression base classifier (with $l2$-regularization) to be the best performing, and these are the results we report on. We implemented our classifier and ran experiments using the Scikit-learn Python library [32].

3.2.3 *Evaluation Methodology*

We tested using two different sizes of the training data for predicting a specific participant's preferences: 75% of 100% of the answers provided by the remaining participants. In all cases, training data also included the participant's own answers to three of the scenarios they were asked about; we tested on the remaining 11 scenarios (10 scenarios in the case of the participants mentioned in Section 3.1).

When predicting comfort level, we report accuracy in two ways, which differ in how they treat predictions when the participant did not have a preference. In the first approach, we counted any prediction as correct if the participant's actual survey response fell in the middle of the Likert scale, i.e., their answer was "Neither Agree nor Disagree." We did this based on the reasoning that if an individual doesn't have an explicit preference, then any prediction would be consistent with that preference. In the second approach, we report accuracy by testing only on scenarios for which a participant did not answer neutrally. This measures how many of a participant's non-neutral preferences can be predicted.

Additionally, for both prediction tasks, we report the results of using a simple majority classifier that classifies each element in the test set as the majority class within the training set.

In each experiment, we randomly selected 50 participants whose answers to predict. We report the accuracy, precision, and recall of the classifier averaged over the 50 participants.

Accuracy is the fraction of predictions that were accurate. Both precision and recall are indicators for measuring the effectiveness of a classifier in predicting positive examples. For predicting comfort, a positive example is a scenario for which the user's answer falls into the "comfortable" category. For predicting allow/deny decisions, a scenario for which a user answers "Allow" is a positive example. *Precision* is the fraction of positive predictions during testing that are actually correct according to the training data. *Recall* is the fraction of all positive examples in the training data that the classifier predicts as positive during testing.

For each participant, we used a form of cross-validation defined as follows:

For $X = 75\%$ or $X = 100\%$ of training data:

- Randomly select 50 participants as targets for prediction.
- For each participant, run 6 different iterations of prediction.
- In each of the 6 iterations, randomly select $X\%$ of training data from the remaining participants and randomly select 3 responses from the total set of scenarios the target was asked

Categories	Tags (Usage)	Examples
Factors (n = 842)	purpose (63%), data (26%), retention (25%), sharing (18%), benefit (17%), location (7%), device (2%),	P880:“It would make me more comfortable knowing where this data was going and how it was going to be used, as well as it being consented.”
Whitelist (n = 350)	safety (42%), anonymous_data (40%), personal_benefit (7%), public (7%), common_good (6%), improve_services (6%)	P908: “If they helped to make me safer in some way.”, P779:“I’d be fine with data that doesn’t identify me.", P121:"That my safety was the reason for it, or saving me money”
Blacklist (n = 474)	biometrics (26%), personal_information (20%), everything (16%), location (13%), private_location (12%), bathroom (9%), video (9%), commercial (8%), government (6%), law_enforcement (5%)	P136:“[..] that they might share the data with other parties [..]. Also, knowing that a retinal or fingerprint scan might be stolen and used to gain access to something else.” P415:“The government spying on me in my home, or private corporations using that data to identify me [..], no way.”
Information (n = 417)	purpose (66%), retention (35%), sharing (21%), collector (15%), access (13%), data_handling (13%), data_security (5%)	P271: “Knowing exactly what the data is used for, where it is stored, who it is shared with, and when it is collected.”
Control (n = 113)	deletion (33%), consent (30%), opt-out (27%), ownership (14%), access (13%), copying (10%)	P913: “Nine times out of ten I won’t care and would be happy to allow it, I just want to be informed and have the ability to deny consent should I choose.”
Risks (n = 298)	misuse (29%), surveillance (18%), data_security (18%), privacy (16%), tracking (12%), intransparency (8%),	P286:“I don’t want my personal information getting into the wrong hands.” P47:“I don’t like the idea of government organizations being alerted of my location at all times.”

Table 3: Categories and codes used to code free text answers. Percentages in brackets are the number of times a code was used when the category was coded, multiple codes could be applied per category. Rows on Factor/Whitelist/Information/Control refer to answer to the question “..what would make you uncomfortable with sharing data in such situations?” Blacklist/Risks stem from the answers to the question about discomfort.

about. This data is used for training; testing is done on the remaining scenarios of the target.

- Calculate the average accuracy, precision, and recall scores averaged over 6 iterations each and over the 50 random participants.

We report on the results of our experiments in Sections 4.2.2 and 4.3.2.

3.3 Qualitative Analysis of Preferences

We also qualitatively analyzed participants’ responses to the free-response questions they were asked at the end of the survey. The answers were coded with regards to five topics: the factors that were mentioned; whether specific scenarios were described as comfortable or uncomfortable; what the participant wants to be informed about; and what means of control (e.g. access, edit, ability to delete) they request. A codebook was developed from 100 answers and applied to another set of 100 answers by two annotators independently. They reached an inter-annotator agreement of 0.89 (Cohen’s Kappa) for whether a topic was addressed and between 0.67 and 0.72 on the actual tags (e.g., which factor was mentioned). After achieving this accuracy, the remaining answers were divided among the two annotators and coded by one annotator each. A summary of categories and codes and their occurrence is shown in Table 3.

4. RESULTS

In this section, we describe our participants and present results regarding participants’ comfort level with different data collection scenarios, their decisions to allow or deny data collection, and desire to be notified.

4.1 Participants

Our survey was completed by 1,014 MTurk workers. We removed the answers of seven participants because they took less than five minutes to complete the survey, while the average completion time was 16 minutes. This resulted in 1,007 participants whose responses we included in our analyses. Participants were required to be from the United States and have a HIT approval rate of above 95%. Table 2 describes participants according to their demographics and privacy concern level. Our participants were slightly better educated and had a higher income than the U.S. average.

4.2 Comfort with Data Collection

In our survey, after presenting each scenario we asked: “How would you feel about the data collection in the situation described above if you were given no additional information about the scenario?” We measured participants’ comfort on a five point Likert scale from “Very Comfortable” to “Very Unfomfortable” with the middle point of “Neither Comfortable Nor Uncomfortable.”

Figure 1 shows the general distribution of participants’ comfort across different levels of each factor. Participants were strongly uncomfortable if the scenarios they were asked about had biometric as data_type (45% strongly uncomfortable), device_type as iris scanner (50% strongly uncomfortable), location as their home (44% strongly uncomfortable), retention as forever (33% strongly uncomfortable), or if other data was inferred from the data collection (42% strongly uncomfortable).

4.2.1 Factors Impacting Comfort Level

Using the best model, we ordered the factors based on their contribution to comfort level by looking at the change in BIC when each factor was added to the null model (the model that has no factor other than random intercept for participants). Table 4 shows the factors ordered by their effect sizes from the most effective factor (the interaction between the data_type and happening_today) to the factor with the lowest effect size (retention). As shown in the table, not all levels of the factors are statistically significant ($p < 0.05$). A positive estimate (effect size) indicates inclination toward comfort and a negative estimate shows inclination toward discomfort.

Scenarios in which video was being collected and participants thought such data collections are happening_today had the greatest positive impact on participant comfort with data collection ($p < 0.05$, coefficient = 1.38). This is in line with our qualitative results, where we found that 38% of all participants mentioned a specific scenario with which they were comfortable (category “whitelist,” Table 3), and from the whitelisted scenarios, 42% mentioned safety, security, or emergency situations as specific purposes for data collection that they would generally approve of. Another 40% of those who whitelisted a scenario were less concerned when anonymous or anonymized data was involved. When an example was given, participants mentioned scenarios involving presence or temperature sensors as ones they would be comfortable with.

Scenarios in which biometric information (e.g., fingerprint, iris image) was being collected and participants thought such data collection is happening_today, had the greatest negative impact on participant comfort ($p < 0.05$, coefficient = 0.89). This is also in line with our qualitative analysis of answers to the question "Keeping in mind the 14 scenarios, what would make you uncomfortable with sharing data in such situations?" In 46% of the answers, participants conveyed one or more specific things that they did not want to happen (coded in category "blacklist," Table 3). Within these answers, the collection of biometric data_type was mentioned by 26%.

Based on previous findings [7], we hypothesized that participants would be less comfortable if a scenario included the explicit notice that collected data would be shared with others (shared). Consistent with that hypothesis, we found that informing participants that data would be shared with third parties (e.g., with the device manufacturer or law enforcement) caused participants to be less comfortable ($p < 0.05$, coefficient = -0.68). The qualitative results show that a minority of participants expressed mistrust of or discomfort with sharing with government (6%) and law enforcement (5%) agencies.

Within the qualitative responses related to discomfort, we also found explanations of why participants did not want to share their data. About 29% of all participants mentioned some perceived risk, ranging from the fear of identity theft or the use of data for other than the stated purpose (misuse) to a general concern about privacy and surveillance in general. Among those that mentioned a perceived risk, 29% feared that their data could be used in a way that would harm them or put them at a disadvantage. About 18% of these answers explicitly mentioned data security issues and leaks as a cause of concern.

> P11: [I'm concerned about] any unique identifiers that could be hacked and then used for identity theft, blackmail, humiliation, etc.

With respect to the location of data collection, most levels had small, positive effect on comfort level. As described above, only scenarios taking place at home had a negative impact on the perceived comfort. Our qualitative results further substantiate this, as participants who mention location as a factor that made them comfortable often cited the dichotomy between public and private places. Data collection in private places is described as highly intrusive while data collection in publicly accessible spaces like libraries or stores was described as "ok." Out of the 474 participants that expressed discomfort with specific scenarios, those that took place in one's home (12%) and in bathrooms (8%) were most frequently mentioned.

The factor retention had the smallest effect size on the results and only short retention times (immediate deletion or storing for a week) had a significant, positive effect on the comfort level. This is in line with the qualitative results were, about 25% of those that mentioned a specific factor in their answers referred to how long their data was stored. Those that explicitly mentioned a time span favored a retention time of less than a week.

4.2.2 Predicting Comfort Level

As explained in Section 3, we trained a machine learning model to predict a participant's comfort based on the significant factors and interactions determined through model selection. The results are shown in Table 5.

The classifier achieved an average accuracy of around 81% over 50 different participants when either 100% or 75% of the other participants' answers are used as training data.

There is a sizable difference in precision and recall depending on whether (1) predictions are counted as correct whenever participants expressed neither a positive nor a negative opinion or (2) scenarios in which participants did not express an opinion are removed from the test data. As per the discussion in Section 3.2.3, both ways of measuring performance are indicative of the utility of using a similar classifier in practice.

Factor	Estimate	Std Err	Z-value	p-value	BIC
data type:happening today baseline=friend's house:not happening today					14633
video:happening today	1.39	0.20	6.83	**0.00**	
biometric:happening today	0.89	0.15	5.80	**0.00**	
presence:happening today	0.91	0.18	12.57	**0.01**	
temperature:happening today	0.95	0.22	4.26	**0.00**	
data (baseline=specific position)					15843
biometric	-1.45	0.13	-11.12	**0.03**	
presence	1.42	0.16	8.99	**0.00**	
temperature	2.50	0.20	12.57	**0.00**	
video	-0.30	0.19	-1.62	0.11	
user perceive benefit:location baseline=beneficial:friend's house					15866
not beneficial:department store	0.00	0.32	0.00	0.99	
purpose unspecified:department store	-0.07	0.24	-0.30	0.76	
not beneficial:house	-0.15	0.48	-0.30	0.76	
purpose unspecified:house	0.05	0.28	0.19	0.85	
not beneficial:library	-0.45	0.33	-1.38	**0.00**	
purpose unspecified:library	-0.17	0.24	-0.70	0.48	
not beneficial:public restroom	-0.40	0.36	-1.10	0.27	
purpose unspecified:public restroom	-0.48	0.26	-1.85	**0.01**	
not beneficial:work	-0.49	0.36	-1.38	0.17	
purpose unspecified:work	-0.11	0.24	-0.47	0.63	
being shared:user perceived benefit					15969
baseline=not being shared:beneficial					
being shared:not beneficial	-0.71	0.19	-3.70	**0.00**	
shared:purpose unspecified	0.37	0.13	2.94	**0.02**	
user perceived benefit (baseline=beneficial)					16055
not beneficial	-1.88	0.34	-5.60	**0.00**	
purpose unspecified	-1.30	0.25	-5.26	**0.04**	
retention:user perceived benefit baseline =unspecific:not beneficial)					16058
not deleted:not beneficial	-0.12	0.22	-0.06	0.96	
purpose specific:not beneficial	-0.30	0.28	-1.08	0.28	
week:not beneficial	0.49	0.23	2.11	**0.00**	
year:not beneficial	0.10	0.24	0.39	0.69	
not deleted:purpose unspecified	-0.43	0.16	-2.69	**0.00**	
week:purpose unspecified	-0.29	0.16	-1.76	0.07	
year:purpose unspecified	-0.22	0.17	-1.31	0.19	
happening within 2 years (baseline=disagree)					16199
agree	0.96	0.11	9.01	**0.00**	
happen today (baseline=disagree)					16491
agree	10.98	333.4	0.03	0.97	
location (baseline=friend's house)					17987
library	1.00	0.18	5.54	**0.00**	
work	0.87	0.18	4.82	**0.01**	
house	-0.88	0.20	-4.34	**0.00**	
department store	0.76	0.18	4.24	**0.00**	
public restroom	0.29	0.19	1.48	0.14	
being shared (baseline=not being shared)					18079
being shared	-0.68	0.09	-7.86	**0.00**	
IUIPC					
collection	-0.59	0.05	-11.47	**0.04**	18081
retention (baseline=not specified)					18103
week	0.25	0.11	2.25	**0.00**	
year	0.16	0.11	1.45	0.14	
purpose specific	0.0.56	0.15	4.85	**0.02**	
not deleted	0.10	0.10	0.99	0.32	

Table 4: Generalized linear mixed model regression output for the comfort level model. A positive estimate (effect size) indicates inclination toward comfort and a negative estimate shows inclination toward discomfort. Factors are ordered by their contribution: the factor with the lowest BIC contributes most to explaining participants' comfort level.

Class.	Training	Neutral	Acc.	Prec.	Recall
ABC	100% (1,006)	correct	81.06%	73.86%	83.06%
ABC	100% (1,006)	excluded	77.53%	54.50%	63.49%
ABC	75% (755)	correct	81.79%	71.30%	78.34%
ABC	75% (755)	excluded	77.67%	54.48%	60.77%
SMC	100% (1,006)	correct	72.03%	71.33%	40.92%
SMC	100% (1,006)	excluded	67.96%	0%	0%

Table 5: Accuracy, precision, and recall of (1) **ABC**: the AdaBoost classifier (with logistic regression as the base learner) and (2) the **SMC**: simple majority classifier, for predicting a user's comfort level with an instance of data collection. "Training" indicates the fraction (and number) of non-test participants used to train the classifier. "Neutral" indicates whether predictions are always counted as correct if a participant didn't indicate a preference for that scenario ("correct") or whether such scenarios are removed from the test set ("excluded").

Class.	Training	Acc.	Prec.	Recall
ABC	100% (1,006 users)	79.09%	76.79%	82.32%
ABC	75% (755 users)	79.09%	76.79%	82.32%
SMC	100% (1,006 users)	52.58%	0%	0%

Table 6: Accuracy, precision, and recall of (1) **ABC**: the AdaBoost classifier (with logistic regression as the base learner) and (2) **SMC**: the simple majority classifier, for predicting a user's decision to allow or deny data collection. "Training" indicates the fraction (and number) of non-test participants used to train the classifier.

Table 5 also describes the performance of our simple majority classifier that uses all non-test participants' answers as training data. These results form a baseline for understanding the performance of the AdaBoost classifier. Although a majority classifier is correct about 70% of the time, AdaBoost additionally correctly predicts more than a third of the predictions that the majority classifier gets wrong.

4.3 Allowing or Denying Data Collection

4.3.1 Factors Impacting Allow/Deny Decisions

We found a set of factors that can explain participants' response to the question: "If you had the choice, would you allow or deny this data collection?" We again ordered factors with respect to their effect size. The interaction of data_type and location has the most impact while shared has the smallest effect. By looking at the coefficient of the levels within each factor we can claim that participants were most likely to deny data collection in scenarios in which their presence was being collected at their workplace. Also, knowing that the data was being shared had the least effect on their preference to deny a data collection. In this model a positive estimate shows likeliness to deny and a negative estimate shows the likeliness to allow a data collection scenario. The regression results are shown in Table 7.

Among the common statistically significant factor levels, the ones that made participants more likely to be comfortable with a data collection also made them more likely to allow the data collection. Many factors were in line between the two models of comfort level and allow/deny such as data_type, location, user_perceived_benefit, shared, retention, happening_today, and within_two_years. However, the best model that described participants' comfort level (Section 4.2) was not the same as the best model that described the desire of participants to allow or deny a data collection. For example, we found that the interaction between data_type and location was the most helpful factor in the allow/deny model,

Factor	Estimate	Std Err	Z-value	p-value	BIC
data:location *baseline=specific position:friend's house*					15232
biometrics:department store	1.58	0.24	6.38	**0.01**	
presence:department store	1.22	0.37	3.3	**0.00**	
temperature:department store	1.61	0.55	2.94	**0.00**	
video: department store	-0.99	0.21	-4.83	**0.00**	
presence: house	0.42	0.41	1.02	0.31	
temperature: house	0.23	0.42	0.54	0.58	
biometrics:library	1.16	0.23	5.01	**0.01**	
presence:library	1.55	0.37	4.1	**0.01**	
temperature:library	1.52	0.43	3.52	**0.00**	
video:library	-0.5	0.2	-2.46	**0.00**	
presence:public restroom	1.87	0.36	5.11	**0.00**	
temperature:public restroom	1.54	0.38	3.99	**0.00**	
video:public restroom	1.36	0.36	3.77	**0.00**	
presence:work	2.11	0.34	6.1	**0.03**	
temperature:work	1.66	0.39	4.29	**0.00**	
being shared:user perceived benefit *baseline=not being shared:beneficial*					15297
being shared:not beneficial	0.62	0.19	3.26	**0.00**	
shared:purpose unspecific	-0.27	0.12	-2.1	**0.04**	
retention:user perceived benefit					15352
not deleted:not beneficial	-0.147	0.226	-0.65	0.515	
purpose-specific:not beneficial	0.39	0.248	1.37	0.17	
week:not beneficial	-0.126	0.24	-0.52	0.6	
year:not beneficial	-0.17	0.24	-0.68	0.49	
not deleted:purpose unspecified	0.45	0.16	2.81	**0.02**	
week:purpose unspecified	0.76	0.16	4.52	**0.00**	
year:purpose unspecified	0.48	0.17	2.85	**0.01**	
user perceived benefit (baseline=beneficial)					15374
not beneficial	2.85	0.17	16.38	**0.00**	
purpose unspecified	1.67	0.17	9.92	**0.01**	
data:happening today *baseline=friend's house:not happening today*					15525
video:happening today	-1.39	0.22	-6.26	**0.00**	
biometric:happening today	-0.78	0.16	-4.89	**0.00**	
presence:happening today	-0.95	0.19	-5.02	**0.02**	
temperature:happening today	-0.9	0.23	-3.87	**0.00**	
happening within 2 years:benefit of scenario *baseline=disagree:benefit to company*					15986
agree: purpose unspecified	0.12	0.36	0.34	0.73	
agree:benefit to user	-0.38	0.23	-1.64	**0.00**	
happening within 2 years (baseline=disagreement)					16751
agreement	-0.72	0.20	-3.7	**0.03**	
data (baseline=specific position)					16872
biometric	0.01	0.24	0.06	0.95	
presence	-2.87	0.35	-8.01	**0.00**	
temperature	-3.66	0.37	-9.66	**0.00**	
video	0.43	0.23	1.82	0.07	
happening today (baseline=disagreement)					17112
agreement	-11.01	349.4	-0.03	0.97	
benefit of scenario (baseline=benefit to company)					18188
benefit to user	-0.46	0.20	-2.30	**0.01**	
purpose unspecified	-1.17	0.27	-4.34	**0.00**	
location (baseline=friend's house)					18569
library	-1.87	0.29	-6.34	**0.02**	
work	-1.96	0.27	-7.34	**0.01**	
house	0.54	0.35	1.52	0.13	
department store	-1.58	0.29	-5.3	**0.00**	
public restroom	-1.23	0.29	-4.17	**0.04**	
retention (baseline=not specified)					18669
week	-0.55	0.11	-4.72	**0.02**	
year	-0.32	0.11	-2.79	**0.00**	
purpose-specific	-0.70	0.12	-5.76	**0.00**	
not deleted	-0.03	0.11	-0.26	0.79	
being shared (baseline=not being shared)					18707
being shared	0.52	0.10	5.41	**0.00**	

Table 7: GLMM Regression Output for the allow-deny model. A positive estimate shows likeliness to deny and a negative estimate shows the likeliness to allow. Factors are ordered by their contribution: the factor with the lowest BIC contributes most to explain participants' desires to allow or deny a data collection.

but this factor was shown to be non-significant in explaining the comfort level. This suggests that being comfortable with a specific data collection instance does not automatically mean that someone would allow it to occur, given the choice.

In the free text answers to the questions about what would make them feel comfortable or uncomfortable with data collection, about 11% of all participants mentioned some type of ability to control collection or use as a requirement for comfort, though our scenarios did not include such a feature. Nevertheless, participants expressed interest in a variety of ways to control their personal information. Within the group that mentioned it, 33% wanted to be granted the ability to delete their data; this would make them feel more comfortable. Another 30% wanted to be asked for consent first, and 27% desired the ability to opt out of the data collection at any time. Multiple participants acknowledged that they would probably not make use of the control options, were they provided.

4.3.2 *Predicting Allow/Deny Decisions*

Using the significant factors and interactions we determined from the model selection, we trained a machine learning model to predict an individual's decision to allow or deny data collection. The results are shown in Table 6. In this experiment, a prediction is made based on the class (allow or deny) that had the higher probability in the prediction. Averaged over 50 test participants, accuracy ranged from 76% to 80% depending on whether we used most (75%) or all of the other participants' data during training.

Table 6 also describes the results of our simple majority classifier when using all other participant's answers as part of the training data. Similar to when predicting comfort, we use the results of this experiment as an intuitive baseline for understanding how well a classifier does if it simply uses the most prevalent preference in the training data.

The average accuracy of the majority classifier of barely over 50% shows that participants' collective preferences were sufficiently evenly split between wanting to allow and deny data collection in general; hence, a classifier that takes more context into account is necessary for effective prediction. The precision and recall values are 0 because the majority class was always to *deny* data collection, resulting in no true positives ever being predicted, which is clearly not representative of an individual's actual preferences.

Understanding how well we can predict an individual's decision to allow or deny data collection is useful in applications such as where a system pre-populates a privacy control panel with an individual's predicted responses. If an individual changes a pre-populated control (i.e., responding with something different than the system's prediction), the system can update its model with this new "correct" answer. Iteratively refining answers until the system is very confident about a decision will ultimately lead—our results suggest—to the majority of answers specific to an individual being predicted with high confidence.

4.4 Data Collection Notification Preferences

We presented participants with questions asking how often they want to be notified about a data collection with three different frequencies. The frequencies are whether they would want to be notified 1) every time, 2) once in a while, or 3) only the first time the data is collected. They were asked to answer their preferences for all three types of notifications on a five point Likert scale ranging from "Strongly Agree" to "Strongly Disagree."

The best models for describing the three frequencies of notifications revealed that participants' preferences for notification changes based on the factors and levels of factors. The three significant factors that were common between all the models were: data_type, location, and the interaction of these two factors. In these models positive coefficients (estimate) show likeliness of participants' desire to get notification about a data collection.

Factor	Estimate	Std Err	Z-value	p-value	BIC
data:user perceived benefit					13467
baseline=friend's house:not beneficial					
biometrics:not beneficial	0.09	0.21	0.46	0.64	
presence:not beneficial	-0.49	0.24	-2.04	**0.00**	
temperature:not beneficial	-0.38	0.35	-1.1	0.27	
video:not beneficial	0.48	0.22	2.19	**0.00**	
biometrics:purpose unspecified	0.88	0.42	2.12	**0.01**	
presence:purpose unspecified	-0.04	0.48	-0.08	0.93	
temperature:purpose unspecified	-0.71	0.46	-1.55	0.12	
video:purpose unspecified	-0.19	0.47	-0.42	0.67	
data:happening within 2 years					13591
baseline = friend's house:disagree					
video:agree	-0.48	0.34	-1.44	0.15	
biometric:agree	-0.01	0.24	-0.04	0.96	
presence:agree	-0.76	0.33	-2.31	**0.02**	
temperature:agree	-0.11	0.39	-2.28	0.78	
being shared:data (baseline = not being shared:specific position)					13738
being shared:data					13738
baseline = not being shared:specific position					
being shared:presence	0.96	0.22	4.39	**0.00**	
being shared:temperature	-0.27	0.2	-1.32	0.18	
being shared:video	0.73	0.17	4.2	**0.01**	
data (baseline = specific position)					14198
biometric	0.17	0.44	0.39	0.7	
presence	-0.57	0.54	-1.07	0.29	
temperature	-1.66	0.54	-3.07	**0.00**	
video	-0.02	0.52	-0.03	0.98	
happening within 2 years (baseline = disagree)					14697
agree	-0.27	0.19	-1.42	0.15	
user perceived benefit (baseline = beneficial)					14923
not beneficial	0.89	0.16	5.45	**0.00**	
purpose unspecified	0.69	0.35	1.94	**0.04**	
benefit of scenario:location					15281
baseline = benefit to company:friend's house					
benefit to user:department store	-0.01	0.25	-0.02	0.98	
purpose unspecified:department store	0.13	0.28	0.46	0.65	
benefit to user:house	-0.65	0.27	-2.38	**0.01**	
purpose unspecified:library	0.71	0.22	3.18	**0.00**	
benefit to user:library	0.31	0.25	1.28	0.2	
benefit to user:public restroom	0.16	0.25	0.62	0.54	
benefit to user:work	0.29	0.24	1.18	0.23	
benefit of scenario (baseline = benefit to company)					15421
benefit to user	-0.26	0.41	-0.66	0.51	
purpose unspecified	-0.77	0.36	-2.12	**0.00**	
location (baseline = friend's house)					15471
library	-1.11	0.19	-5.58	**0.01**	
work	-1.09	0.19	-5.57	**0.00**	
house	0.79	0.21	3.81	**0.00**	
department store	-0.69	0.20	-3.41	**0.03**	
public restroom	-0.29	0.19	1.48	0.14	
being shared (baseline = not being shared)					15539
being shared	0.17	0.11	1.62	0.11	

Table 8: Generalized Linear Mixed Model Regression output for every-time notification. A positive coefficient (estimate) shows likeliness of participants' desire to get notification about a data collection every time. Factors are ordered by their contribution: the factor with the lowest BIC contributes most to explain participants' preferences about every-time notification.

In the free text answers, 41% of all participants mentioned that being informed would help them feel comfortable, indicated by phrases like "I would want to know..." or "If they would tell me...". Within that group, purpose, a factor heavily dependent on data_type and location, was mentioned by the majority (66%) as something that they would want to be informed about. It was followed by retention (35%), a factor not found in the model. 15% also explicitly requested information on who would be collecting the data (code "collector"). In addition, 13% of this group wanted to be informed about who is accessing the data and 5% want to be informed about steps taken to ensure the security of the collected data. Eight percent of the participants showed some kind of mistrust related to the purpose of data collection described in the scenarios. This was expressed in various ways, from demanding to know "exactly" what was stored and requesting "guarantees" to asking for honesty or expressing general concern about their privacy.

> P928: I like honesty, and with companies being honest and open about why they are sharing data, it makes it a lot easier for me to be comfortable.

More detailed information was also requested about potential risks and how their data was protected against misuse.

4.4.1 Notification Every Time

We measured participants' preferences to get notified about a type of data collection every time it occurred by their answers to the question "I would want my mobile phone to notify me every time this data collection occurs." The factors in the order of their size of effect are shown in Table 8. The most effective factor in explaining participants' desire to be notified every time was the interaction between data_type and user_perceived_benefit, while the factor that had the smallest effect size was shared. Looking at the levels of these factors, it seems that participants were most likely to want to be notified every time when their biometrics were being collected for an unspecified purpose. Also, knowing that the data was being shared had the least effect on participants' desire to be notified every time the data collection occurred.

4.4.2 Notification Once in a While

We measured participants' preferences to being notified only once in a while about a type of data collection by their answers to the question "I would want my mobile phone to notify me every once in a while when this data collection occurs." The results in the order of effect size are shown in Table 9. The model selection algorithm showed that the most effective factor in explaining participants' desire to be notified once in a while was data_type and the least effective factor was the interaction between data_type and location. The coefficients of the levels within these factors show that participants were most likely to want to be notified every once in a while when their biometric was being collected and their desire to get notification every once in a while was least effected by knowing that their presence was being collected while they were at a department store.

4.4.3 Notification the First Time

We measured participants' preferences to being notified only the first time about a type of data collection by their answers to the question, "I would want my mobile phone to notify me only the first time this data collection occurs." Table 10 shows the factors we got from the model selection in order of the effect size. The most effective factor in explaining participants' desire to be notified for the first time was user_perceived_benefit and the factor with the smallest effect size was the interaction between the data_type and location. More specifically, participants were most likely to want to get a notification only the first time if the data collection was not beneficial to them. Also their desire to get notified only for the first time was least effected when their biometric was being collected while they were at a department store.

Factor	Estimate	Std Err	Z-value	p-value	BIC
data (baseline = specific position)					14172
biometric	-0.56	0.16	-3.35	**0.00**	
presence	-0.07	0.24	-0.27	0.78	
temperature	-0.03	0.25	-0.13	0.9	
video	-0.42	0.14	-3.07	**0.01**	
IUIPC					
control	-0.29	0.07	-4.03	**0.00**	14231
location (baseline = friend's house)					14238
library	0.48	0.22	2.21	**0.02**	
work	0.64	0.18	3.63	**0.00**	
house	0.31	0.19	1.63	0.1	
department store	0.29	0.22	1.36	0.18	
public restroom	0.26	0.22	1.19	0.23	
data:location					14243
baseline=specific position:friend's house					
biometric:department store	0.24	0.21	1.14	0.26	
biometric:library	-0.02	0.2	-0.09	0.92	
presence:department store	-0.62	0.29	-2.14	**0.00**	
presence:home	-0.001	0.27	-0.006	0.99	
presence:library	-0.85	0.29	-2.83	**0.00**	
presence:public restroom	-0.67	0.29	-2.29	**0.03**	
presence:work	-0.48	0.25	-1.87	0.61	
temperature:department store	-0.76	0.38	-1.98	**0.00**	
temperature:home	0.52	0.28	1.86	0.62	
temperature:library	-1.34	0.33	-4.06	**0.00**	
temperature:public restroom	-0.86	0.31	-2.87	**0.00**	
temperature:work	-0.87	0.28	-3.12	**0.04**	
video:department store	-0.09	0.19	-0.48	0.62	
video:library	-0.11	0.19	-0.54	0.59	
video:public restroom	-0.30	0.25	-1.20	0.22	

Table 9: Generalized Linear Mixed Model Regression output for once-in-a-while notification. A positive coefficient (estimate) shows likeliness of participants' desire to get notification about a data collection every once in a while. Factors are ordered by their contribution: the factor with the lowest BIC contributes most to explain participants' preferences for once-in-a-while notification.

4.4.4 Summary of Data Collection

At the end of each survey, we asked participants the question "Keeping in mind the 14 scenarios, how often would you be interested in seeing a summary of all such data collection?" Participants could select either every day, every month, every year, or never. Answers varied, with 23% (n = 232) saying they would like a daily summary and 63% (633) selecting a monthly summary. Additionally, 8% (85) would have liked a summary every year and 6% (57) never wanted to receive one.

5. LIMITATIONS

Our study has limitations common to many user studies and to user studies in the area of privacy. Although the demographic attributes of the participant group are, except for the reported income, close to the US average, Mechanical Turk workers do not reflect the general population. Prior research has shown that Mechanical Turk workers are more privacy-sensitive than the general population [16]. It has also been has shown that self reports about privacy preferences often differ from actual behavior. This is referred to as the "privacy paradox" [10, 1]. Our study may be susceptible to this bias because the scenarios were abstract and participants were asked to imagine themselves in situations they may not have encountered. In addition,

Factor	Estimate	Std Err	Z-value	p-value	BIC
user perceived benefit (baseline=beneficial)					14487
not beneficial	-0.47	0.07	-7.09	**0.01**	
purpose unspecified	-0.32	0.05	-6.08	**0.00**	
location (baseline=friend's house)					14567
library	0.74	0.22	3.37	**0.02**	
work	0.86	0.18	4.76	**0.00**	
house	0.08	0.19	0.41	0.68	
department store	0.75	0.22	3.36	**0.03**	
public restroom	0.61	0.22	2.81	**0.00**	
data (baseline=specific position)					14587
biometric	0.17	0.17	1.02	0.31	
presence	0.78	0.24	3.24	**0.00**	
temperature	0.81	0.25	3.30	**0.00**	
video	0.00	0.13	-0.02	0.99	
data:location *baseline = specific position:friend's house*					14617
biometric:department store	-0.58	0.21	-2.79	**0.00**	
biometric:library	-0.30	0.2	-1.51	0.13	
presence:department store	-1.05	0.29	-3.66	**0.00**	
presence:home	-0.23	0.27	-0.83	0.41	
presence:library	-1.19	0.29	-4.02	**0.02**	
presence:public restroom	-1.19	0.29	-4.13	**0.00**	
presence:work	-0.48	0.25	-1.86	0.06	
temperature:department store	-1.61	0.38	-4.26	**0.00**	
temperature:home	0.23	0.28	0.82	0.41	
temperature:library	-1.35	0.32	-4.18	**0.00**	
temperature:public restroom	-1.09	0.31	-3.58	**0.00**	
temperature:work	-1.17	0.28	-4.19	**0.01**	
video:department store	-0.16	0.19	-0.85	0.39	
video:library	-0.17	0.19	-0.89	0.37	
video:public restroom	-0.54	0.25	-1.20	0.22	

Table 10: Generalized Linear Mixed Model Regression output for first-time-only notification. A positive coefficient (estimate) shows likeliness of participants' desire to get notification about a data collection only the first time. Factors are ordered by their contribution: the factor with the lowest BIC contributes most to explain participants' preferences for first-time-only notification.

some of the scenarios in our study were designed to be realistic based on common data collection and use practices that are happening

today, while others were designed to be more forward-looking. We decided to have some less-realistic scenarios because we hypothesized that there is a relation between participants' comfort level about each vignette and their perception of how realistic it is. Nevertheless, participants may have been asked about situations which they are not typically put in, influencing their decisions.

Despite these limitations, presenting a large variety of scenarios to participants allowed us to explore situations that do not currently happen but may be similar to situations that will happen in the future. Since the Internet of Things is still an emerging field, it is not possible to describe situations that are realistic to all participants who may never have had an IoT device or never have faced a situation in which an IoT sensor is collecting data.

6. DISCUSSION

Our results demonstrate varied privacy concerns, both across IoT scenarios and across participants. Our results also indicate that participants are more comfortable about data collection when classical privacy and data protection rules, such as the Fair Information Practices, are applied and individuals are given an explanation about why their data is being collected. However, other results underline the need for technology to support the awareness of data collection and that can meet the different desires for being notified.

6.1 Privacy Preferences Are Complex

How individuals feel about different data collection scenarios depends on various things. Individual preference play as much a role as social norms and expectations.

On one hand, our analyses show that participants are largely in agreement on a number of practices where social norms are in place that define what is acceptable and what is not. For example, participants expressed more comfort with data collection in public spaces, but rejected scenarios that described video cameras used in private rooms and shared with law enforcement. This is likely related to a long, western tradition of public/private dichotomy. However, this dichotomy is challenged by smart-home technology with centralized, cloud-based services that do not follow expectation of "what happens at home stays at home." For example, Samsung received criticism for advising the public not to have private conversations in front of their smart TV [14] as it uses a third party speech-to-text service for voice commands. Smart-home device manufacturers should be aware and respectful of individuals' mental models of data collection within the home and do their best to communicate practices that may be surprising to their customers.

On the other hand, we saw a large number of scenarios in which there was no clear indication of what is generally acceptable. For example, participants showed a high variance in the level of comfort with respect to the collection and storage of movement patterns at their workplace for the purpose of optimizing heating and cooling. Social norms have yet to emerge with respect to technology that has just recently become available. However, scenarios like these also reflect how individual preferences might differ in the long run. Individuals have to weigh their potential loss of privacy, due to camera surveillance against the benefit of reduced energy consumption. The complexity of this individual decision process is also reflected by the fact that our models describing the comfort level and the choice to allow or deny a data collection do not completely overlap. Here individual concerns about what might happen to the data, in combination with personal experience (e.g., how much one trusts her employer), play a role in determining whether or not one feels comfortable with the data collection and will allow it.

6.2 Addressing Privacy Concerns

Both the qualitative and quantitative data show that participants prefer anonymous data collection. Temperature and presence sensors produce data that are not immediately identifying and participants consistently expressed higher comfort with these scenarios. This finding was further reinforced by our free-text results, as anonymous data was the second most mentioned preference for data collection. This is further confirmed through interviews done in a previous study [7]. The relatively high discomfort with data inference, combined with high comfort regarding collection of anonymous data indicates that people may be generally unaware that with the Internet of Things it will be easier to re-identify individuals from otherwise anonymous data. In light of our findings, it is likely that this is something that would cause discomfort. This gap in understanding should be kept in mind when providing privacy information for IoT data collection.

We found that participants favor short retention times and are more comfortable when data is deleted after its purpose is met, or not kept longer than a week. Insights from the free-text responses indicate that this is related to an increased awareness of data breaches, the fear of misuse of data, and concerns regarding bad data security practices at companies. As previous research has shown, a growing number of people have already experienced misuse of their data [34]. With the growing number of IoT devices, the probability of data breaches further increases, resulting in higher concern and less trust in the technology. To address these types of concerns, IoT device manufacturers should take precautions, both technical and administrative, to protect their customers' data and communicate these practices to the public.

6.3 Towards Awareness and Control

Approaches for eliciting consent or providing information are less likely to work in the IoT setting. For example, a classic privacy policy cannot be shown on many types of IoT devices, such as a smart watch. Still, people demand information about the entity collecting data, the purpose of the collection, the benefit they receive from it, and the retention period of the collected data.

In open-ended responses, participants explicitly asked for transparency in data collection and its handling. Discomfort increases when data is shared with third parties or used to infer additional information. Participants want to be informed not only about the purpose of data collection and the handling of data, but also possible security risks associated. This finding is also confirmed by previous work which found through interviews that transparency about the data collected and the purpose of the collection influence comfort levels for data collection by IoT devices [7].

Additionally, our results show that how often and about what participants want to be informed is greatly dependent on individual comfort levels. But information requests also heavily depend on whether or not individuals think a use of their data is beneficial to them or serves a greater good. To answer this question even semi-automatically requires more specific and neutral information about the purpose of a data collection. We also saw that two thirds of participants would appreciate a monthly summary about what data has been collected about them (see section 4.4).

To develop technical support for this is a major challenge in a fractured IoT landscape that still lacks standardization. One option to streamline these efforts, at least on a smaller scale like in smart homes, would be to build upon the Manufacture Usage Description Specification [11] to include information on purposes of data collection and simplify the aggregation of information about data collection.

Our analysis suggests that many people want to retain control of their personal data. Future IoT services should take this into consideration when designing privacy notices instead of creating more "one-size fits all" policies.

More specifically, we suggest the adoption of the idea of personalized privacy assistants (PPA) already used in the context of mobile apps [25]. A PPA may be a tool or agent running on behalf of each individual that can proactively predict their decision to allow or deny data collection, relieving the individual of making decisions when they can be predicted with high accuracy. This predictive model could be used to, i.e., pre-populate a privacy control panel with individuals' preferences. In a deployed system, we could use a form of online machine learning to continue to update the model to a specific individual's preferences. Our predictive model 4.3 showed that with a few data points per individual (three), we could predict the rest of their eleven answers with an average accuracy of 88%. In a deployed system, we expect the model would have more specific data points about individuals on which to base predictions, which would be even more accurate.

7. CONCLUSIONS

In this paper we reported on a large-scale vignette study on privacy concerns related to the Internet of Things. We asked 1,007 participants to rate realistic scenarios about data collection occurring in multiple contexts. Our results enhance the findings of previous, mostly qualitative research with statistical evidence that identifies specific factors that impact individuals' privacy concerns. Among these factors are the type of data that is collected, retention time, third-party sharing, perceived benefit, and the location at which an IoT device collects data. The statistical results are confirmed by analyses of the free-text responses, which emphasize concerns regarding the collection of biometric data as well as data collection occurring in private spaces.

Based on our findings, we made recommendations for designing IoT services and applications. People favor data collection in which they cannot be identified immediately. They also do not want inferences to be made from otherwise anonymous data. We found that participants want to be informed about various details of data collection, such as what the data is used for and how long it will be stored.

8. ACKNOWLEDGEMENTS

This research has been supported in part by DARPA and the Air Force Research Laboratory under agreement number FA8750-15-2-0277 and by the National Science Foundation under grant SBE-1513957. The US Government is authorized to reproduce and distribute reprints for Governmental purposes not withstanding any copyright notation thereon. Additional support has also been provided by Google. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA, the Air Force Research Laboratory, the NSF, Google, or the US Government.

9. REFERENCES

[1] Alessandro Acquisti and Ralph Gross. 2006. Imagined communities: Awareness, information sharing, and privacy on the Facebook. In *Proc. PETS*.

[2] Ivor D. Addo, Sheikh Iqbal Ahamed, Stephen S. Yau, and Arun Balaji Buduru. 2014. A Reference Architecture for Improving Security and Privacy in Internet of Things Applications. In *IEEE Third International Conference on Mobile Services*. 108–115. DOI: http://dx.doi.org/10.1109/MobServ.2014.24

[3] Christiane Atzmüller and Peter M. Steiner. 2010. Experimental vignette studies in survey research. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences* 6, 3 (2010), 128–138. DOI: http://dx.doi.org/10.1027/1614-2241/a000014

[4] Debjanee Barua, Judy Kay, and Cécile Paris. 2013. Viewing and Controlling Personal Sensor Data: What Do Users Want? In *Persuasive Technology*, Shlomo Berkovsky and Jill Freyne (Eds.). Number 7822 in Lecture Notes in Computer Science. Springer Berlin Heidelberg, 15–26. http://link.springer.com/chapter/10.1007/978-3-642-37157-8_4 DOI: 10.1007/978-3-642-37157-8_4.

[5] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67, 1 (2015), 1–48. DOI: http://dx.doi.org/10.18637/jss.v067.i01

[6] Pankaj Bhaskar and Sheikh Iqbal Ahamed. 2007. Privacy in Pervasive Computing and Open Issues. In *Proceedings of the The Second International Conference on Availability, Reliability and Security, ARES 2007, The International Dependability Conference - Bridging Theory and Practice*. 147–154. DOI: http://dx.doi.org/10.1109/ARES.2007.115

[7] Igor Bilogrevic and Martin Ortlieb. 2016. "If You Put All The Pieces Together...": Attitudes Towards Data Combination and Sharing Across Services and Companies. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. DOI: http://dx.doi.org/10.1145/2858036.2858432

[8] Richard Chow, Serge Egelman, Raghudeep Kannavara, Hosub Lee, Suyash Misra, and Edward Wang. 2015. HCI in Business: A collaboration with academia in IoT privacy. In *International Conference on HCI in Business*. Springer, 679–687.

[9] The Federal Trade Commission. 2015. *Internet of Things: Privacy & Security in a Connected World*. Technical Report. Federal Trade Commission. Accessed Mar. 2017.

[10] Catherine Dwyer, Starr Roxanne Hiltz, and Katia Passerini. 2007. Trust and privacy concern within social networking sites: A comparison of Facebook and MySpace. In *Proc. AMCIS*.

[11] E. Lear, R. Droms, and D. Romascanu. 2017. *Manufacturer Usage Description Specification*. Internet-Draft draft-ietf-opsawg-mud-04. IETF Network Working Group. https://datatracker.ietf.org/doc/draft-ietf-opsawg-mud/?include_text=1

[12] Serge Egelman, Raghudeep Kannavara, and Richard Chow. 2015. Is this thing on?: Crowdsourcing privacy indicators for ubiquitous sensing platforms. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1669–1678.

[13] Janet Finch. 1987. The vignette technique in survey research. *Sociology* (1987), 105–114.

[14] David Goldman. 2015. Your Samsung TV is eavesdropping on your private conversations. (Feb. 2015). http://money.cnn.com/2015/02/09/technology/security/samsung-smart-tv-privacy/index.html

[15] Joseph B Kadane and Nicole A Lazar. 2004. Methods and criteria for model selection. *Journal of the American statistical Association* 99, 465 (2004), 279–290.

[16] Ruogu Kang, Stephanie Brown, Laura Dabbish, and Sara B Kiesler. 2014. Privacy Attitudes of Mechanical Turk Workers and the US Public. In *SOUPS*. 37–49.

[17] Predrag Klasnja, Sunny Consolvo, Tanzeem Choudhury, Richard Beckwith, and Jeffrey Hightower. 2009. Exploring Privacy Concerns about Personal Sensing. In *Pervasive Computing*, Hideyuki Tokuda, Michael Beigl, Adrian Friday, A. J. Bernheim Brush, and Yoshito Tobe (Eds.). Number 5538 in Lecture Notes in Computer Science. Springer Berlin Heidelberg, 176–183. http://link.springer.com/chapter/10.1007/978-3-642-01516-8_13 DOI: 10.1007/978-3-642-01516-8_13.

[18] Scott Lederer, Jason I Hong, Anind K Dey, and James A Landay. 2004. Personal privacy through understanding and action: five pitfalls for designers. *Personal and Ubiquitous Computing* 8, 6 (2004), 440–454.

[19] Scott Lederer, Jennifer Mankoff, and Anind K. Dey. 2003. Who wants to know what when? Privacy preference determinants in ubiquitous computing. In *Extended abstracts of the 2003 Conference on Human Factors in Computing Systems, CHI 2003, Ft. Lauderdale, Florida, USA, April 5-10, 2003*. 724–725. DOI: http://dx.doi.org/10.1145/765891.765952

[20] Hosub Lee and Alfred Kobsa. 2016. Understanding User Privacy in Internet of Things Environments. *Internet of Things (WF-IoT)* (2016).

[21] Hosub Lee and Alfred Kobsa. 2017. Privacy Preference Modeling and Prediction in a Simulated Campuswide IoT Environment. In *Proceedings of the 15th IEEE Conference on Pervasive Computing and Communications*. IEEE.

[22] Pedro Giovanni Leon, Blase Ur, Yang Wang, Manya Sleeper, Rebecca Balebako, Richard Shay, Lujo Bauer, Mihai Christodorescu, and Lorrie Faith Cranor. 2013. What matters to users?: factors that affect users' willingness to share information with online advertisers. In *Proceedings of the ninth symposium on usable privacy and security*. ACM, 7.

[23] Jialiu Lin, Shahriyar Amini, Jason I Hong, Norman Sadeh, Janne Lindqvist, and Joy Zhang. 2012. Expectation and purpose: understanding users' mental models of mobile app privacy through crowdsourcing. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. ACM, 501–510.

[24] Jialiu Lin, Bin Liu, Norman Sadeh, and Jason I Hong. 2014. Modeling usersĂŹ mobile app privacy preferences: Restoring usability in a sea of permission settings. In *Symposium on Usable Privacy and Security (SOUPS)*, Vol. 40.

[25] Bin Liu, Mads Schaarup Andersen, Florian Schaub, Hazim Almuhimedi, Shikun Aerin Zhang, Norman Sadeh, Yuvraj Agarwal, and Alessandro Acquisti. 2016. Follow My Recommendations: A Personalized Assistant for Mobile App Permissions. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*.

[26] Bin Liu, Jialiu Lin, and Norman Sadeh. 2014. Reconciling mobile app privacy and usability on smartphones: Could user privacy profiles help?. In *Proceedings of the 23rd international conference on World wide web*. ACM, 201–212.

[27] Naresh K Malhotra, Sung S Kim, and James Agarwal. 2004. Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. *Information systems research* 15, 4 (2004), 336–355.

[28] Kirsten E Martin. 2012. Diminished or just different? A factorial vignette study of privacy as a social contract. *Journal of Business Ethics* 111, 4 (2012), 519–539.

[29] Kirsten E Martin and Helen Nissenbaum. 2016. Measuring Privacy: An Empirical Test Using Context To Expose Confounding Variables. *Columbia Science and Technology Law Review* 18 (2016), 176–218.

[30] Sharad Mehrotra, Alfred Kobsa, Nalini Venkatasubramanian, and Siva Raj Rajagopalan. 2016. TIPPERS: A privacy cognizant IoT environment. In *Pervasive Computing and Communication Workshops (PerCom Workshops), 2016 IEEE International Conference on*. IEEE, 1–6.

[31] Helen Nissenbaum. 2009. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press.

[32] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[33] Charith Perera, Rajiv Ranjan, Lizhe Wang, Samee Ullah Khan, and Albert Y. Zomaya. 2015. Big Data Privacy in the Internet of Things Era. *IT Professional* 17, 3 (2015), 32–39. DOI:http://dx.doi.org/10.1109/MITP.2015.34

[34] Lee Rainie, Sara Kiesler, Ruogu Kang, and Mary Madden. 2013. Anonymity, Privacy, and Security Online. (Sept. 2013). http://www.pewinternet.org/2013/09/05/anonymity-privacy-and-security-online/

[35] Beate Rössler. 2005. *The value of privacy* (english ed ed.). Polity, Cambridge, UK ; Malden, MA.

[36] Norman Sadeh, Jason Hong, Lorrie Cranor, Ian Fette, Patrick Kelley, Madhu Prabaker, and Jinghai Rao. 2009. Understanding and capturing people's privacy policies in a mobile social networking application. *Personal and Ubiquitous Computing* 13, 6 (2009), 401–412.

[37] Janice Y Tsai, Patrick Kelley, Paul Drielsma, Lorrie Faith Cranor, Jason Hong, and Norman Sadeh. 2009. Who's viewed you?: the impact of feedback in a mobile location-sharing application. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2003–2012.

[38] Arijit Ukil, Soma Bandyopadhyay, and Arpan Pal. 2015. Privacy for IoT: Involuntary privacy enablement for smart energy systems. In *2015 IEEE International Conference on Communications, ICC 2015, London, United Kingdom, June 8-12, 2015*. 536–541. DOI: http://dx.doi.org/10.1109/ICC.2015.7248377

APPENDIX

Appendix material is formatted differently than what appeared in the survey seen by participants.

A. SAMPLE SURVEY SCENARIO

You are at a **friend's house**. All rooms have **presence sensors that are used to determine when to switch on and off the lights to reduce costs and save energy**. You are **not told how long the data will be kept.**

Q1. This use of my data would be beneficial to me. (Answered on a five point Likert scale from "Strongly Agree" to "Strongly Disagree")

Q2. I think scenarios like this happen today. (Answered on a five point Likert scale from "Strongly Agree" to "Strongly Disagree")

Q3. (If "disagree" or "strongly disagree" for Q2) I think scenarios like this will happen within 2 years. (Answered on a five point Likert scale from "Strongly Agree" to "Strongly Disagree")

Q4. (If "disagree" or "strongly disagree" for Q3) I think scenarios like this will happen within 10 years. (Answered on a five point Likert scale from "Strongly Agree" to "Strongly Disagree")

Q5a. How would you feel about the data collection in the situation described above if you were not told with whom the data would be shared, how long it would be kept or how long it would be used for? (Answered on a five point Likert scale from "Very Comfortable" to "Very Uncomfortable")

Q5b. How would you feel about the data collection in the situation described above if you were given no additional information about the scenario? (Answered on a five point Likert scale from "Very Comfortable" to "Very Uncomfortable")

Q6a. I would want my mobile phone to notify me every time this data collection occurs. (Answered on a five point Likert scale from "Strongly Agree" to "Strongly Disagree")

Q6b. I would want my mobile phone to notify me only the first time this data collection occurs. (Answered on a five point Likert scale from "Strongly Agree" to "Strongly Disagree")

Q6c. I would want my mobile phone to notify me every once in a while when this data collection occurs. (Answered on a five point Likert scale from "Strongly Agree" to "Strongly Disagree")

Q7. If you had the choice, would you allow or deny this data collection? (Choices: Allow, Deny)

B. SUMMARY QUESTIONS

Q1. Keeping in mind the 14 scenarios, how often would you be interested in seeing a summary of all such data collection? (Choices: Every day, Every month, Every year, Never)

Q2. Keeping in mind the 14 scenarios, what would make you comfortable with sharing data in such situations?

Q3. Keeping in mind the 14 scenarios, what would make you uncomfortable with sharing data in such situations?

C. IUIPC QUESTIONS

Participants answered the following questions on a seven point Likert scale from "Strongly Aagree" to "Strongly Disagree"

1. Consumer online privacy is really a matter of consumers' right to exercise control and autonomy over decisions about how their information is collected, used, and shared.
2. Consumer control of personal information lies at the heart of consumer privacy.
3. I believe that online privacy is invaded when control is lost or unwillingly reduced as a result of a marketing transaction.
4. Companies seeking information online should disclose the way the data are collected, processed, and used.
5. A good consumer online privacy policy should have a clear and conspicuous disclosure. It is very important to me that I am aware and knowledgeable about how my personal information will be used.
6. It usually bothers me when online companies ask me for personal information.
7. When online companies ask me for personal information, I sometimes think twice before providing it.
8. It bothers me to give personal information to so many online companies.
9. I'm concerned that online companies are collecting too much personal information about me.

D. DEMOGRAPHIC QUESTIONS

Q1. How old are you?

Q2. What is your gender? (Choices: Female, Male, Other, Prefer not to answer)

Q3. What is the highest degree you have earned? (Choices: No high school degree, High school degree, College degree, Professional degree (masters/PhD), Associates degree, Medical degree, Prefer not to answer)

Q4. What is your income range? (Choices: Less than $15,000/ year, $15,000/ year - $24,999/year, $25,000/ year - $34,999/ year, $35,000/ year - $49,999/ year, $50,000/ year - $74,999/ year, $75,000/ year - $99,999/ year, $100,000/ year - $149,999/year, $150,000/year - $199,999/ year, $200,000/ year and above, Prefer not to answer)