Natã M. Barbosa*, Joon S. Park, Yaxing Yao, and Yang Wang

"What if?" Predicting Individual Users' Smart Home Privacy Preferences and Their Changes

Abstract: Smart home devices challenge a long-held notion that the home is a private and protected place. With this in mind, many developers market their products with a focus on privacy in order to gain user trust, yet privacy tensions arise with the growing adoption of these devices and the risk of inappropriate data practices in the smart home (e.g., secondary use of collected data). Therefore, it is important for developers to consider individual user preferences and how they would change under varying circumstances, in order to identify actionable steps towards developing user trust and exercising privacy-preserving data practices. To help achieve this, we present the design and evaluation of machine learning models that predict (1) personalized allow/deny decisions for different information flows involving various attributes, purposes, and devices (AUC .868), (2) what circumstances may change original decisions (AUC .899), and (3) how much (US dollars) one may be willing to pay or receive in exchange for smart home privacy (RMSE 12.459). We show how developers can use our models to derive actionable steps toward privacy-preserving data practices in the smart home.

Keywords: smart home, internet of things, privacy preferences, machine learning, economics of privacy

DOI Editor to enter DOI
Received ..; revised ..; accepted ...

1 Introduction

The promise of technology that can make life easier inside one's home drives the concept of the smart home. Smart home devices include thermostats, door locks, and a variety of other devices that can be remotely controlled and are Internet-connected, with some even listening all the time (e.g., voice-activated assistants). The ubiquitous adoption of these devices challenges the centuries-old notion that the home is a private, protected, and intimate place, and generates user privacy concerns that stand in the way of widespread adoption of smart home devices, especially because they can enable monitoring in otherwise personal spaces [18, 27].

However, experts believe that although consumers are worried about privacy, optimism bias will outweigh privacy concerns and ultimately drive adoption of these devices [31]. Optimism bias refers to underestimating the chances of being subject to a negative event. For example, people may think that they will not be a target of security attacks or privacy breaches, therefore engaging in unsafe practices such as reusing passwords and not adopting security tools such as two-factor authentication. Smart home device adoption is indeed rising: in 2017 alone, 20 million Amazon Echo devices and 7 million Google Home devices were sold [8]. These numbers are expected to continue growing significantly as smart home devices offer more features and convenience at a low cost. A recent report by Juniper Research predicts that stand-alone voice assistants are expected to be part of 55% of US households by 2022, with a number of installed devices to achieve 175 million [17]. Such widespread adoption of smart home devices could critically change entrenched norms around information flows [28] and create opportunities for large-scale appropriation of data collected inside one's home, thus threatening the privacy of individuals in unprecedented ways.

Many developers are being wary and marketing their products with privacy in mind, but privacy tensions remain around data practices in the smart home. Such tensions are justified by the potential for abuse, misuse, and appropriation of user data. For example, while developers may say they will not sell user data, nothing stops them from changing their policy in the future and using data collected originally for automation, to serve unforeseen, secondary purposes (e.g., targeted advertising). In fact, lead developers already use such data for that purpose, such as Google with the Google Home Mini. Such data practices puts developers at an advantage, being able to commoditize user data and amplify their knowledge about one's life inside their homes, creating "sequels" to Web and smartphone tracking.

***Corresponding Author: Natã M. Barbosa:** Syracuse University, E-mail: nmbarbos@syr.edu
Joon S. Park: Syracuse University, E-mail: jspark@syr.edu
Yaxing Yao: Syracuse University, E-mail: yyao08@syr.edu
Yang Wang: Syracuse University, E-mail: ywang@syr.edu

From a theoretical angle, according to Nissenbaum's contextual integrity theory [28], activities that violate information gathering and dissemination norms expected in a given context, and that cross the governing norms of distribution within it, constitute a privacy violation. Contextual integrity theory also holds that notions of privacy also rely on ethical concerns that arise over time. Therefore, one could argue that if such extensive appropriation is to take place without a regard to the privacy norms appropriate to the context (i.e., people's homes), then they will pose serious consequences that violate societal principles and values in regards to privacy, ultimately resulting in the spanning of long-settled boundaries [30]. Therefore, preventing improper information flows becomes a necessity in order to alleviate privacy tensions, develop user trust toward smart home devices [41], and ultimately, protect the home's long-held privacy norms at a societal level [41].

To address this need, we present machine learning models that can be used by developers to derive actionable steps toward respecting the privacy of users in a personalized way. We developed three models from survey data: (1) a model that can predict allow/deny preferences based on one's current information privacy inclinations as well as personal and home attributes, purposes of use, and devices that may be involved in a given smart home information flow (AUC .868); (2) a model that can predict, for each information flow, what circumstances can change users' original preferences (e.g., what if data are [not] transmitted securely? what if data are [not] collected frequently?) (AUC .899) and (3) a model that can be used for predicting how much (in US dollars) users would be willing to pay extra for added privacy protections or accept as a discount/refund to allow collection/sharing of data (RMSE 12.459). By using these models, smart home developers can not only obtain fine-grained, personalized user preferences on a large scale, but also identify potentially inappropriate data practices based on such preferences and unveil actionable steps in order to respect the privacy of users.

2 Background and Related Work

In collecting smart home privacy preferences via a survey, automatically predicting preferences and their changes, and learning the privacy value in the smart home, we situate our research in light of prior works on privacy preferences in the Internet of Things (IoT), modeling of privacy preferences, and privacy valuations.

2.1 Privacy Preferences in the IoT

Broad IoT. Because the IoT has the potential to significantly increase sensing capabilities, user concerns about privacy are commonplace. This has motivated researchers to look into user preferences regarding data collection in IoT environments on a large scale. Notable studies of this nature are Lee and Kobsa's [20] and Naeini *et al.*'s [27]. These studies were conducted as combinatorial scenario-based surveys with the goal of identifying the impact of different contextual factors on privacy decisions, such as where data collection takes place, what data types (e.g., video, photo) are involved, who collects the data, reason for data collection, and the retention period following the data collection. These studies revealed that privacy preferences vary greatly based on the data types (e.g., video, biometrics), purpose of use, entities (e.g., government), whether information flows are used for safety, convenience, and the benefit of the user, and whether the data are inferred or not. Both studies also indicate that users are mostly uncomfortable with information flows at private places such as the home. These studies are foundational in regards to privacy preferences and expectations of users in the broad IoT and provide avenues for further research.

Unlike these studies, our survey focuses exclusively on the smart home, and explores different contextual factors that may influence privacy decisions more deeply, such as users' Internet privacy concerns, comfort levels toward manufacturer, third parties, and government, personal and home attributes (rather than data types like video, image or voice), purposes of use, and situational factors known to change privacy decisions in other domains (e.g., the Web, smartphones). By "zooming into" the smart home, we evaluate preferences in light of the impregnability of the home [28] and the potential for secondary use. We evaluate personal and home attributes because they are directly associated with knowledge of activities inside the home, which users are uncomfortable with [7]. In addition, the tracking of device attributes and events may also lead to physical safety vulnerabilities [9], so we include device states, actions, and events as attributes in our scenarios. This way we can know, for example, if users would be comfortable with the use of personal, device, and home attributes for the purpose of targeted advertising.

Smart Home. A number of previous studies have focused on specific devices in the smart home environment, hinting at different factors that may affect people's privacy perceptions towards them. For example, there have been studies about smart home technolo-

gies for elders [10], assistive technologies [13], smart toys [25], and smart home devices in general [41]. Such prior case studies hinted that certain factors can cause user privacy concerns more than others, and that users are overall uncomfortable with potential monitoring of their activities inside the home. While these studies focused on specific use cases such as smart toys and assistive technology for older adults, they reveal one interesting discrepancy about users' choices involving the benefits that smart home technologies can bring and the privacy issues raised by adopting such technologies. For example, some choose privacy over benefits [13], while others' needs will outweigh any privacy concerns [11], with trust toward the manufacturer being important (e.g., [41]).

Apthorpe *et al.* [5] proposed a combinatorial method to obtain privacy norms in the smart home based on the contextual integrity privacy framework. Their methodology is very similar to Lee and Kobsa's [20] and Naeini's [27], and ultimately most similar to our own survey's methodology (i.e., the survey we conducted to collect our training data), given its highly contextual approach to capturing privacy norms in the smart home. They conducted a study which revealed that people may be uncomfortable with entities other than the manufacturer accessing smart home data, and that consent and the ability to use the data for emergencies contribute the most toward increased comfort. On the other hand, targeted advertising and permanent storage contributed the most toward discomfort.

In conducting a similar survey, we followed their recommendations moving forward, such as considering data practices in the smartphone domain and their transfer into the smart home, and considering attributes associated with other people at the home (e.g., guests). The authors, however, did not attempt to model personalized privacy preferences with machine learning, and they did not consider pre-established Internet privacy concerns of users (e.g., IUIPC [24]) in their data collection methodology, which we do. In addition, the major differences from Apthorpe *et al.*'s survey to ours are that we cover a more comprehensive list of attributes, purposes of use, and situational factors that can influence privacy decisions identified previously in the smartphone [19, 21, 38] and online behavioral advertising domains [26, 37]. These situational factors are especially relevant in the smart home because IoT devices vary greatly in their sensing capabilities as well as the entities behind them, for example, involving companies of many sizes and backgrounds which may provide different levels of stability, security, and reliability [15].

2.2 Modeling Privacy Preferences

Other Domains. With the growing number of devices, apps, and resources users have to manage daily, protecting individual privacy can be challenging and burdensome. For this reason, recent works in other domains, especially mobile phone app permissions, have proposed effective ways to use machine learning to assist users in managing privacy preferences (e.g., [22, 29, 39, 40]). These works have focused on developing and evaluating tools that predict preferences about app permissions for the user, essentially making a decision on their behalf. Such studies also indicate that through a small number of questions, a large number of preferences can be accurately obtained, effectively reducing user burden. Authors also advocate for models to account for purposes of use, uncertainty, contextual factors, and the malleability of privacy preferences (e.g., avoid one-shot, binary decisions). More importantly, as pointed by Olejnik *et al.* [29], in developing future automated privacy management, it is important to identify what data flows are likely to "defy" users' expectations in a given context. We agree, and further argue that asking the types of *"what if?"* questions like we do in our work enables developers to not only capture the malleability of privacy preferences, but also gain user trust; an important factor in the smart home. In addition, Liu *et al.* [22] assert that privacy assistants can be used in domains where privacy configuration is an issue, with one of such domains being the IoT, where devices lack contextual cues, for instance, having small screens or no screens at all.

While these research efforts were conducted in different domains, their findings indicate that modeling privacy preferences is a promising approach, and many relevant implications can be learned from them. One common aspect of these works is that they use behavior data to train machine learning models, whereas we use data capturing expectations/attitudes. This poses a limitation to our models in that they may learn from expectations and attitudes rather than behavior, and traditionally, these are known to differ. This discrepancy is also known as the privacy paradox, which reflects a deviation between attitudes and behavior when it comes to privacy decisions. Nonetheless, these works have also suggested that data from expectations have been effective in making privacy recommendations, and that such models could be adjusted as behavior data are collected over time [22, 40]. We note, however, that the limited contextual cues in IoT devices make it difficult to gather behavior data (e.g., no or small screens [22]), in addition to the fact that, at the time of this writing, there is

no established permission model for the IoT. From these prior works on modeling privacy preferences, we learned that automated decisions for privacy management can reduce privacy violations, but such decisions must consider context and the malleability of privacy decisions. Our machine learning pipeline incorporates these aspects, and in a new domain: the smart home. There have been attempts, however, to model privacy settings in the IoT more broadly, which we describe next.

IoT. Regarding modeling privacy preferences in the IoT, the closest work to our own is Bahirat *et al.*'s [6], where authors evaluated models based on Lee and Kobsa's survey data [20], which capture user preferences in the IoT, in order to create user-facing privacy-preserving profiles. Combining clustering algorithms with decision tree models for classification, the authors reached 82% accuracy in predicting user preferences that could be used to derive IoT users' privacy settings. The authors used three variables about each scenario to represent a user's attitude: risk, comfort, and appropriateness, averaged across 14 scenarios, and created decision trees with entities and data types.

Besides offering increased performance in preference prediction (5% increase) and being focused more specifically on smart home rather than the IoT more broadly, our work offers a number of meaningful advantages compared to Bahirat *et al.*'s. First, our approach enables more convenient collection of user features (four scenarios *vs.* 14). Second, by using validated privacy constructs such as the Internet Users' Information Privacy Concerns (IUIPC) scale [24], we also capture each user's existing attitudes/concerns toward online privacy. Third, one of our models enables the prediction of potential changes in comfort levels. Last, but not least, one of our models enables the prediction of monetary value associated with privacy in the smart home.

While our work offers these advantages, Bahirat *et al.*'s decision-tree approach is more interpretable – a requirement for user-facing tools – whereas our models can be used by developers in the back-end on a large scale, not to strictly enforce default settings, but to assist in identifying acceptable data practices for their users.

2.3 Value of Privacy

One aspect of privacy decision-making is that it often involves cost-benefit assessments from users. In fact, some argue that this relationship between privacy, costs, and expected benefits, warrants approaching privacy from an economics perspective (e.g., [1]). In this aspect, the smart home domain has two interesting, distinct characteristics compared to other domains: (1) devices are paid for, rather than free-to-use and (2) they are inserted into a privacy-by-default environment i.e. the home. The former puts a price tag on the expected added convenience such devices can offer, and the latter poses a risk, thus making for an interesting research case.

Studies of privacy valuations often measure the dollar value associated with privacy by involving choices of either selling/disclosing and protecting personal information, for example, by offering discounts and/or extra protective features in a "willing-to-accept" and "willing-to-protect" manner (e.g., [4, 14]). These studies have shown that people would accept more money to disclose personal information than they would be willing to pay to protect the very same information. They also indicate that people give more value to privacy when they already have it than when they do not (i.e., they are loss-averse). Previous studies have also looked into interdependent privacy valuations, such as the privacy value of a friend's data (e.g., [32–34]). Such studies show that people will put more value on their own privacy than the privacy of others, but that people's assessments also depend on context, such as whether information from their friends is really necessary.

In our scenario-based survey, we adopt the willingness-to-accept and willingness-to-protect approach to learn people's privacy values associated with the purchase of a smart home voice assistant costing $49 (USD), and later evaluate whether such dollar amounts can be predicted using machine learning algorithms.

3 Method

Our goal was to create machine learning models that could predict personalized, appropriate privacy preferences about different information flows, as well as identify circumstances that could change such preferences. For example, would the user allow the use of their indoor location for targeted advertising? *What if* consent is obtained for doing so? How much would a user pay for extra privacy protections? Such models could be used by a developer to exercise privacy-preserving data practices and gain user trust in the smart home. Developers can reproduce our steps to collect training data from their user base and create similar models of their own. To obtain such data, we conducted a scenario-based factorial online survey with participants based in the US, which we present in the following section.

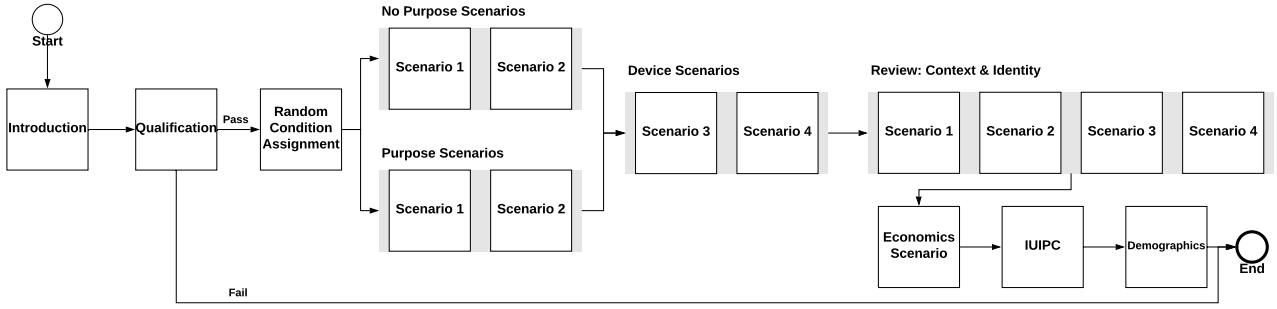

Fig. 1. Survey Workflow. Participants responded and reviewed four randomly generated scenarios representing information flows, in addition to one economics-related scenario involving a voice assistant, IUIPC questions, and demographics.

3.1 Survey

3.1.1 Design

We conducted a factorial vignette study similar in design and structure to Naeini *et al.*'s [27] and Apthorpe *et al.*'s [5]. We used Amazon Mechanical Turk (AMT) to conduct the survey with US-based participants. The survey consisted of asking users for comfort levels, allow/deny choice, and notification frequency for four randomly generated information flows with combinations of attributes, purposes, and devices. A full list of these components can be found in appendix Table 3, and the rationale behind their choice can be found in the appendix Section A.1.1. In addition to collecting original preferences, participants were asked to review situational factors, inspired by prior works, that could make them more or less comfortable with the scenarios.

Besides responding to each information flow, we asked participants a scenario-based economics question about a voice assistant, asked them to complete the Internet Users' Information Privacy Concerns (IUIPC) questionnaire [24], and collected demographics. In the economics scenario, participants had to enter how much they would be willing to pay extra for added privacy protections, or take as a discount in the purchase (or refund) to allow the manufacturer to collect and share their data involving a $49 (USD) voice assistant.

Participants were asked to explain every response provided to the survey through mandatory open-ended text fields (e.g., answers for each information flow, situational factors, and economics-related scenario). Such responses were used to check the data provided by respondents for quality and consistency. We manually checked each open-ended response, and removed all data from 25 participants whose answers were not meaningful and/or seemed like random copy-paste. We also included an open-ended, mandatory field at the end of the survey to gather feedback from participants about things that were unclear or any issues that they faced throughout the survey. Although we closely monitored each response to this question, participants did not raise any concerns. In fact, many praised the quality of our experiment through such field. We manually provided bonuses for participants who took longer than average to finish.

In the end, our survey collected data from 698 participants, with a median time to complete of 19 minutes. Participants were compensated with $1.50 (USD) for taking the survey. We recruited participants from the United States with over 95% approval rate on previously submitted work in the AMT platform. As an additional qualification step, participants were introduced to smart home devices based on the Wikipedia definition[1] and a photo of a smart home device (a smart thermostat). Then, they were asked to select three devices they thought to be a smart home device, from photos of six devices including a DSLR camera, a desk lamp, a blender, a smart thermostat, a voice assistant, and a smart bulb. Only participants who selected the three smart home devices were allowed to complete our survey. In the end, we collected preferences for 2,792 scenarios, with four scenarios answered per participant.

3.1.2 Survey Workflow

In this section, we provide details about our survey workflow, summarized in Figure 1.

1. Scenarios. First, participants were presented with the four scenarios, one at a time. Two of the scenarios had a random device involved for which the attribute was device-related (e.g., device states, device actions,

1 https://en.wikipedia.org/wiki/Home_automation

device status). For half of the participants, assigned randomly, the purpose in the information flow was deliberately omitted, since in realistic settings the purpose of data collection is not always clear, which can make users uncomfortable. Below is a scenario example:

"The manufacturer/developer of your smart home device is accessing or inferring ***[age of people at home]****, for example,* ***[the age of all the people who visit and live in your home]****. They are using this information for* ***[company revenue]****, for example,* ***[for the profit of a company who is behind your smart device (e.g., manufacturer, retailer, etc.)]****."*

Depending on the information flow/scenario, the content inside square brackets would change to indicate the attribute, attribute description, purpose, and purpose description, exactly as in Table 3 (appendix). Following the presentation of the vignette, participants were asked to provide their level of comfort with the given scenario, on a scale of 1 to 5 (i.e., very uncomfortable to very comfortable). Then, participants were asked to also provide levels of comfort in the same scale should the manufacturer share the data with a third party, and with the government (one comfort level for each). Following comfort levels, participants were asked to indicate whether if given the choice, they would allow or deny the information flow, and how frequently they would like to be notified about it: "never," "only the first time," "once in a while," or "always."

2. Review. Then, after collecting comfort levels, choice, and notification preferences for each scenario, we asked each participant to review a number of situational factors and check which ones would make them more comfortable (if originally indicated uncomfortable i.e., 1-2) or less comfortable (if they originally indicated comfortable i.e., 3-5) with the scenario. These situational factors were inspired by prior works in the domains of online behavioral advertising and smartphones and are known to affect people's privacy preferences toward information flows (e.g., if manufacturer is well known, if user can benefit from data collection, etc.). Participants were asked to select at most three (enforced) from a list of 13 factors (Table 3 in appendix). In addition, participants were asked to give a comfort level, in the same scale, if their identity (e.g., their name, address, or other identifiable information) were used in the original information flow, given that this was known to cause discomfort from previous studies (e.g., [27]).

3. Economics. For the economics-related scenario, each participant was randomly assigned to one of the four conditions: (1) purchasing a voice assistant and paying extra for added privacy protections; (2) purchasing a voice assistant and getting a discount for fewer privacy protections; (3) owning a voice assistant and paying a one-time fee for added privacy protections and (4) owning a voice assistant and accepting a refund for fewer privacy protections. For all four conditions, participants were first introduced to a voice assistant with a photo of Amazon's Alexa and a brief description, where no indication of the brand/entity behind the product was given. Each scenario posed the voice assistant as costing $49 (USD). In this portion of the survey, added privacy was described as *"more privacy controls and protections such as limited collection and sharing of your personal information,"* and fewer privacy protections was implicitly expressed as *"allowing the manufacturer to collect and share personal information."* We did not explicitly frame the questions with "fewer privacy protections" not to prime participants. Instead, "fewer privacy protections" was implicitly captured by the combined framing of the scenarios and questions, in the following format:

*"Consider a scenario where you [****are looking to purchase a voice assistant that costs OR had a voice assistant for which you paid****] $49. The voice assistant [****has OR has little to no****] privacy controls and protections against collection and sharing of your personal information"*

where the content inside square brackets is defined based on each participant's scenario condition.

After the scenario was presented, participants were then asked to enter the amount corresponding to the question in their condition (i.e., pay extra when purchasing, paying one-time fee after purchase, discount when purchasing, refund after purchase). For example:

"How much would you be willing to take as a discount off the price tag in exchange for allowing the manufacturer to collect and share personal information in the future? Please specify the amount in dollars (number entry)" (example for condition 2: discount at purchase)

4. IUIPC. Following the economics-related scenario, we asked participants to answer IUIPC questions regarding *Awareness*, *Collection*, and *Control* [24]. These allowed us to gauge the level of Internet privacy concerns of our participants, as well as to have machine learning features representing existing privacy concerns.

5. Demographics. Finally, we asked participants for demographic information such as gender, age bracket, whether they owned a smart home device, time spent on the Internet weekly, income bracket, marital status, household size, and whether they had children. Table 4 (appendix) shows the demographics of participants – 49% of survey participants indicated they already owned a smart home device.

3.2 Machine Learning

Below we describe in detail our machine learning experiments, including goals, data collection and preparation, feature engineering, and model selection and evaluation.

Overall Goals. With data from our survey, we were first interested in predicting *allow* or *deny* (binary) preferences given a user's stated privacy attitudes (IUIPC), attributes, purposes, devices, and comfort levels involved in different information flows. Second, we wanted to be able to identify which factors – for a given information flow – would change the original preference of a user toward being *more* or *less* comfortable (binary). Third, we wanted to predict how much (numerical) a user would be willing to pay extra or accept in exchange for *"more"* or *"fewer"* privacy controls either at or after the time of purchase. More importantly, we wanted to be able to model user preferences without requiring too much information from the user. That is, we were interested in knowing how accurately a model would be able to predict allow/deny preferences, factors that could change their preferences, and privacy value in US Dollars, with minimum user effort. This means a potential user of our models would complete the IUIPC questionnaire and give their comfort levels for four randomly generated scenarios – the same number of scenarios in our survey – from which we would take the average comfort levels to use as features, and generate combinations of scenarios with the different attributes, purposes, and devices to serve as personalized preferences. We would also be able to identify, using combinatorics, which factors – when present – would change original preferences, and how much one would be willing to pay extra/accept for added/fewer privacy controls.

Pros and Cons. The main advantage of our approach is the ability for developers to gather potential preferences for a large number of scenarios and distill actionable steps from them. Informally, it is a way to ask *"what if?"* questions in order to understand what may or may not be appropriate for a particular user. Our approach can also be used by developers to align their practices with privacy expectations of their users on a large scale. For example, by using our pipeline, developers can quickly identify whether a new data practice (e.g., use age of people at home for home automation) would be considered appropriate by their user base, that is, what percentage of users would or would not allow the new practice, and what can be done from the developer's standpoint that can make users more comfortable about it. When communicated correctly, such a practice could also increase consumer trust and drive adoption.

A major disadvantage of our approach, however, is that it is based on stated attitudes, which are known to deviate from actual behavior when it comes to privacy decisions. This deviation can limit the use of our models for other applications, such as creating default profiles. Our approach also introduces a new risk, given the ability for developers to identify ways to "profit" from data practices with the least resistance from their users, that is, by indirectly asking questions such as *"what secondary uses would be appropriate according to my user base?"* While we argue that this would not constitute a privacy violation if data practices are aligned with individual preferences, developers can still control how such scenario-based questions are asked, which can prime users for biased preferences. How to address this priming issue deserves further research, which could lead to the development of standardized constructs.

Data Sets. From our survey responses, we produced two data sets: one data set containing preferences and another containing the dollar amounts given to the economics-related scenario. The first data set consists of 2,792 rows containing user responses to the different information flows presented in the survey. That is, each row indicates an attribute, purpose, and device, along with the given comfort level (1-5) for the manufacturer, third party, government, and if identity is included, allow/deny preference, and notification frequency (1-4) indicated by the respondent. Each respondent produced four of such rows and is identified by a randomly generated code indicated in the row. Each row also indicates which situational factors were selected for the scenario, and toward which direction (e.g., "more" or "less" comfort), represented via a presence-absence matrix, where each factor is a column, and "1" is indicated if the factor was selected or "0" otherwise. Because each participant responded to only one economics-related scenario, our second data set contains 698 rows; one row per participant. This data set indicates, in each row, the dollar amount provided by the respondent, and the condition to which the respondent was randomly assigned (1-4). In order to prepare these data sets for our experiments, we conducted the data preparation and feature engineering steps described in appendix Section A.2.

Feature Selection. We selected our features based on features known to affect privacy preferences from prior works: because users have different levels of privacy concerns, we considered IUIPC values; because privacy concerns are sensitive to the entities behind the data collection, we considered the comfort levels toward different entities (e.g., manufacturer, third party, government); because users make privacy-for-convenience

trade-offs, we considered the average notification frequency selected by participants as a proxy for "how much they would like to be bothered."

Models. We created three different models.

The first model is a classification model used to predict allow/deny preferences for different information flows (binary dependent variable), given different attributes, purposes of use, and devices involved in an information flow, as well as privacy attitudes of users such as comfort levels towards different entities (e.g., manufacturer, government, third party) and their attitude toward online privacy (i.e., IUIPC values).

The second model is a classification model used to predict "comfort change" (i.e., "more" or "less" comfortable, binary dependent variable) given different information flows and "selected" or "not selected" situational factors. Such a model can be used to identify situational factors that can make users either more or less comfortable with a given information flow. The value for the dependent variable was determined automatically during the survey, where the question posed was to select situational factors that would make participants either less or more comfortable, depending on their original comfort level toward the manufacturer (e.g., more comfortable if uncomfortable, less comfortable otherwise). For this model, in addition to the dependent variables used for predicting allow/deny choices (e.g., attributes, purposes), we used each situational factor with a value of 0 (unchecked) or 1 (checked) as features (i.e., a presence-absence matrix). Predictions with this model can then be made by observing the model predictions for n choose k factors (e.g., $k = 3$ for the survey) that can change a user's original comfort levels upward or downward. That is, by using a trained model with 13 choose 3 combinations of situational factors as independent variables, one can identify which factors can change comfort levels toward a given information flow.

The third model is a regression model predicting the dollar amount (numerical dependent variable) that users would pay for added or fewer privacy protections either at or after purchase time of a voice assistant, given IUIPC values, comfort levels, and notification frequency. The first and second models use the scenario data set with 2,792 rows (four per respondent), and the third model uses the economics data set with 698 rows.

Model Selection and Evaluation. We iteratively evaluated increasingly complex models by adding features, one at a time, starting with only the IUIPC values (Awareness, Collection, and Control), then adding average comfort toward manufacturer, third party, government, identity, and notification frequency, one at a time. We also experimented with cluster values based on iteratively adding these features in the clustering process *in lieu* of using the raw values. We used the Area Under the Receiving Operating Characteristics (AUC-ROC) curve, also known as AUC, as the performance metric for classification of (1) allow/deny decisions and (2) less or more comfort – the ROC curve shows the performance of a classifier in regards to the True Positive Rate (TPR) and False Positive Rate (FPR). For regression of the dollar value, we used the Root Mean Squared Error (RMSE) as the performance metric.

We used two libraries to conduct our experiments: PySpark's MLlib and scikit-learn. Accordingly, we report the results of our experiments with both libraries. While both libraries are open source, the advantage of using PySpark is that it demonstrates how developers could use our approach with big data to predict preferences and their changes for potentially millions of users. The advantage of using scikit-learn is that it is the most popular library for machine learning, and it does not require a lot of infrastructure to replicate our experiments. We used each library's implementations of Logistic Regression, Decision Tree, Naive Bayes (NB), Random Forest (RF), Multilayer Perceptron (MLP) with three hidden layers with 64 units each, and Gradient Boosting Trees (GBT) for classification. For regression, we used Linear Regression, Decision Tree regression, Random Forest regression, and Gradient Boosting Tree regression. We randomly split our data into 60/30/10, that is, 60% used for training, 30% for validation, and 10% for testing. In splitting the data, we ensured there was no "cross-presence" of scenarios from the same participant in the different splits. In other words, the splits were made by participant IDs first, then the scenarios were selected based on participant IDs. We picked the best models based on the performance on the validation split, but we also report results of 10-fold cross validation for each model. We also estimate the generalization performance of our models by evaluating them with a hold-out test split (10%). Finally, we interpret the learned coefficients of the allow/deny model.

Minor performance differences between the two libraries are observed due to the different default parameters (e.g., number of steps, solvers, regularization) used in the algorithms of each library, and one would have to match the parameters in order to obtain the exact same coefficients and performance. We opted to use the default parameters in each library for simplicity and reproducibility, so there were minor differences. In places where we do not report both libraries' numbers (e.g., Abstract, Conclusion), we report the lower performance.

4 Results

We first summarize the results of our survey in order to familiarize the reader with our data. Then, we present the results of our machine learning experiments.

4.1 Survey

Our survey findings suggest that people may be most comfortable with information flows for which the purpose is home automation, control, and safety – that is, primary, intended purposes of smart home devices. On the other hand, attributes linked to demographics and that hint at habits can cause the most discomfort. Participants in which the economics scenario indicated they already had privacy protections valued it more than those in scenarios where they did not have them, thus confirming loss aversion. We present more details below.

4.1.1 Allow/Deny

As a summary, Figure 2 shows the percentage of "Deny" decisions by attribute and purpose. In general, comfort levels participants provided in the survey were in line with allow/deny preferences. Overall, participants were uncomfortable with most information flows, with the median comfort level toward the manufacturer being 2: "somewhat uncomfortable."

Attributes. Overall, participants would mostly deny information flows with attributes involving demographics such as age and gender, as well as attributes that would enable monitoring such as communications, destinations, indoor location, habits and lifestyle, and number of people at home. On the other hand, participants would mostly allow attributes not directly associated with themselves, such as device information, weather, energy use, and outside or home temperature.
Purposes. From the purposes of use included in the survey, participants would mostly deny – regardless of attributes – purposes such as identity linking, legal actions, price discrimination, targeted ads, and user tracking. Participants would mostly allow information flows for which the purpose is home safety, customized experiences, home control, and home automation – that is, intended purposes of smart home devices. It is very clear from these findings that smart home users may be very uncomfortable with data being used for purposes beyond the convenience which smart home devices intend to offer (e.g., home control, automation, safety).

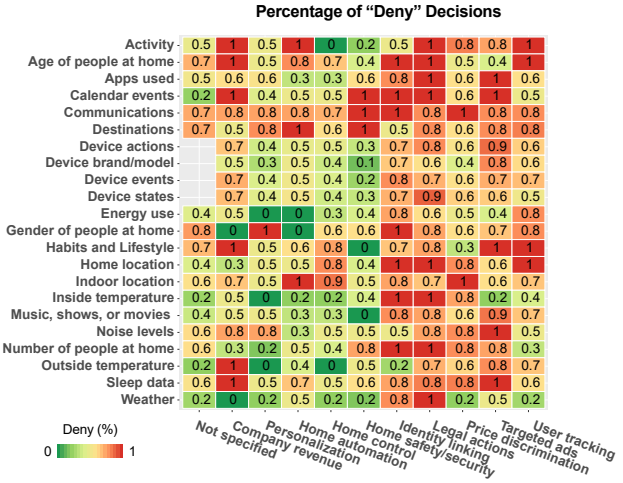

Fig. 2. Percentage of "Deny" decisions for information flows involving different attributes and purposes of use. Demographics and activities were mostly denied as well as secondary purposes.

Finally, the average notification frequency for each attribute and purpose combination reflects comfort levels and preferences in participants' responses. For example, participants wanted to be notified more often for attributes involving demographics and habits, as well as for purposes such as identity linking, legal actions, price discrimination, and user tracking.

Entities. In general, participants indicated being more comfortable with the manufacturer entity involved in the information flow (Mean=2.61, SD=1.38, Mdn=2), followed by the government (Mean=1.94, SD=1.12, Mdn=2), and third parties (Mean=1.82, SD=1.14, Mdn=1). Participants' given average level of comfort when identity is involved in the information flow was 1.94 (SD=1.19, Mdn=1). While it was expected that the level of comfort would go down when user identity is involved, it was surprising to see that in general, participants were more comfortable with the government than with third parties.

4.1.2 Situational Factors

The top situational factors to make participants more comfortable were: if consent was given (48.5%), if they could control when data were shared (38.08%), if the data involved were not sensitive (30.51%), if they were aware of when the information flow occurred (28.49%), and if the data were handled securely (22.26%).

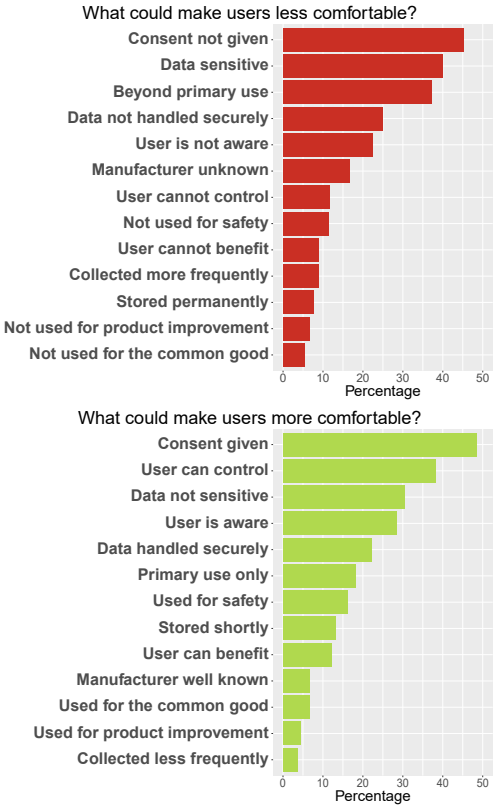

Fig. 3. Top: Percentage of responses for which the situation was selected as making the participant less comfortable. Bottom: Percentage of responses for which the situation was selected that would make the participant more comfortable.

The most chosen factors that could make participants less comfortable were: if consent was not given (45.08%), if it involved sensitive information (39.85%), if the data were used beyond primary purposes (37.16%), if data were not handled securely (24.98%), and if participants were not aware of when the information flows occurred (22.37%). These percentages indicate the number of scenarios where the situational factor was chosen.

Figure 3 shows the percentages of selected situational factors making users less (top) or more (bottom) comfortable with the information flows. By identifying these situational factors in each information flow, developers could take actionable steps to make users more comfortable with certain information flows, while avoiding practices that would make users less comfortable. For example, when asking about an information flow that a user would deny, the manufacturer could notify the user that while they expect the user to deny it, it is only going to be used for the primary purpose of automating the home. Similarly, user trust could be gained if users were aware of when certain information flows occur and if they are given the ability to control them.

4.1.3 Economics of Privacy

The average dollar amount specified by participants as an extra amount they would pay for added privacy protections when purchasing a $49 voice assistant was $14.4 (Mdn=$15, SD=$12), while the average amount for a one-time fee for added privacy protections was $13.3 (Mdn=$10, SD=$13.4).

The average amount of the discount participants would be willing to take when purchasing the voice assistant to allow data collection and sharing was $12 (Mdn=$0, SD=$17.9), and $11.8 (Mdn=$0, SD=$18.9) as the average amount given by participants indicating the one-time refund they would be willing to take to allow the manufacturer to collect and share data.

Interestingly, 53.5% of participants indicated that they would not be willing to take any discount in exchange for their privacy, providing a $0 response. Similarly, 46.2% of participants in the scenario where they owned a voice assistant and were offered a refund for "decreased privacy" responded with a $0 for a refund they would be willing to take in exchange for their privacy. Contrasting these $0 responses with the other conditions, 17% of participants in the "pay extra when purchasing" provided a $0 response, while 24.5% provided $0 as a one-time fee they would be willing to pay for added privacy protections. We have asked each participant to explain their answer and we did a thematic analysis of their explanations for $0 responses. This analysis revealed that 62% of participants in scenarios where privacy protections were already included said that they would not be willing to "put a price on" their privacy. In addition, 58% of participants in the scenario without

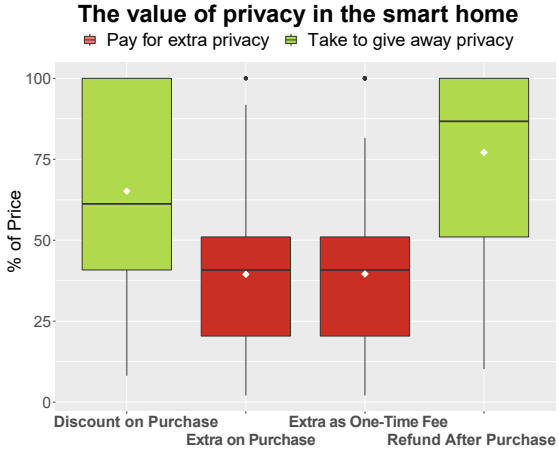

Fig. 4. Distribution of dollar amounts given in each economic scenario as a percentage of $49 (price of voice assistant), with $0 responses removed. White markers indicate average.

privacy protections said that they would expect privacy to be already included and therefore they would not pay extra for it. Among $0 respondents, 59.6% indicated not owning a smart home device.

We also looked at extreme amounts provided by participants. From participants in the "discount when purchasing" condition, 9.9% of participants entered the full price of the voice assistant ($49) as the discount, with the same number of participants indicating $49 for a potential refund in exchange for their privacy.

Given the many $0 responses indicating unwillingness to pay or accept, we removed the $0 responses, and then the average amount indicated as the discount participants would be willing to take when purchasing in exchange for their privacy was $31.9 (Mdn=$30, SD=$14.7), and $37.8 (Mdn=$42.5, SD=$13) for a refund. On the other hand, participants gave an average of $19.3 (Mdn=$20, SD=$9.95) as the extra amount they would pay for privacy protections when purchasing the device, and $19.4 (Mdn=$20, SD=$12) as the one-time fee they would be willing to pay for privacy protections.

Figure 4 shows the distribution of dollar amounts given in each economics scenario, excluding $0 responses. These findings suggest that potential smart home users would value privacy protections more when they already have them than when they do not, corroborating loss aversion observed in previous works [2].

Purpose of Use	Choice	# Actual	# Classified
Company revenue	Allow	4	5
	Deny	10	9
Personalization	Allow	6	7
	Deny	6	5
Home automation	Allow	7	8
	Deny	8	7
Home control	Allow	13	12
	Deny	4	5
Home safety/security	Allow	9	8
	Deny	6	7
Identity linking	Allow	5	4
	Deny	10	11
Legal actions	Allow	5	4
	Deny	12	13
Price discrimination	Allow	7	5
	Deny	8	10
Targeted ads	Allow	4	2
	Deny	10	12
User tracking	Allow	5	6
	Deny	17	16

Table 1. Allow/Deny classification of information flows involving different purposes in the test set.

4.2 Machine Learning Predictions

In this section, we present the results of our machine learning experiments to predict allow/deny choices, predict comfort changes under the presence of different situational factors, and predict the dollar amount associated with privacy in the smart home in a scenario involving a $49 voice assistant. Table 6 (appendix) shows detailed steps during model selection and their results, where we iteratively added features to the models and observed the AUC and RMSE.

4.2.1 Predict Allow/Deny

Model Selection. For allow/deny decisions, the best performance observed in the model selection process using PySpark was a logistic regression model having the following features: IUIPC scores, attribute, purpose, device, and average comfort levels for each entity, namely manufacturer, third party, and government, with a validation AUC of .861. With scikit-learn, the best model also had a validation AUC of .861 and included the same features, except the average comfort for third party and government. To evaluate sensitivity to data set splits, we performed 10-fold cross validation with scikit-learn, which resulted in a .859 AUC. The best-performing model was selected based on its performance on the validation set. While we report performance with both libraries, we will use scikit-learn models for complementary reports in this section (i.e., test set examples, examples of use) due to the library being easier to use.

In selecting a model, inspired by the work of Bahirat *et al.* [6], we also experimented with clustering participants (K-means) based on the IUIPC scores, average comfort levels, and average notification frequency. We experimented with five to three clusters with the same combination of features described in Table 6 (appendix), that is, we clustered based on those features and used the cluster as a feature, along with attributes, purposes, and devices. The best-performing model with this clustering approach was a logistic regression with PySpark, whose resulting validation AUC was .7593 (.7245 for the test set). Due to the lower performance, we did not evaluate models with clusters any further and did not include them in Table 6. We suspect that by clustering participants, information is lost (e.g., variance in the different IUIPC constructs and comfort levels, average over four scenarios used in our work rather than 14).

Model Evaluation. To measure the generalization performance of the selected model, we used the hold-out test set of 10% of rows resulting from the data set splits. Predictions on the test set resulted in an AUC of .868 and .871 on the test set with PySpark and scikit-learn, respectively. Table 1 shows examples of predictions made with scikit-learn on the test set for different purposes. False positives (19 out of 44 mistakes) were mostly given for device attributes (e.g., actions, events, brand/model), but for purposes that would be mostly denied, such as user tracking and company revenue.

Model Interpretation. We interpret this model by looking at the coefficients learned from the logistic regression from both libraries. We had to add regularization to PySpark's model as a fine-tuning step because the resulting coefficients were too large (e.g., 4), so we suspected overfitting. When we added elastic net regularization (*regParam*=0.02, *elasticNetParam*=0.2), this was resolved at a minor cost to performance (.858 validation AUC and .861 test AUC). Scikit-learn had regularization by default. Both libraries converged on very similar models, with top 10 largest coefficients in each direction (deny or allow) being roughly the same (Table 5, appendix). For PySpark, the top five coefficients contributing toward "Deny" were Legal Actions (-1.026), Communications (-.960), Identity Linking (-.952), Age of people (-.798), and Targeted Ads (-.727). For scikit-learn, these were Communications (-1.251), Legal Actions (-1.139), Age of People (-1.128), Identity Linking (-.828) and Targeted Ads (-.760). For coefficients contributing toward "Allow," PySpark's model had Comfort Manufacturer (1.237), Outside Temperature (.787), Weather (.737), Inside Temperature (.668), and Personalization (.632). Similarly, scikit-learn had Comfort Manufacturer (6.346), Inside Temperature (1.191), Weather (1.030), Home Safety (.943), and Outside Temperature (.932). Table 5 (appendix) shows the list of top 10 largest coefficients each way, for both libraries. Different coefficients between both libraries are due to default parameters, and so are the minor performance differences. For example, PySpark's logistic regression uses elastic net regularization and mini-batch stochastic gradient descent, while scikit-learn's uses L2 regularization by default, with LBFGS solver (version 0.22). In order to obtain the same coefficients and models, these parameters would have to be matched. Nonetheless, we used the default implementations for simplicity and reproducibility. From the coefficients, we observe that secondary purposes of use such as legal actions and targeted ads contribute heavily toward a "Deny" decision, as well as personal attributes hinting at demographics and habits. On the other hand, comfort toward the manufacturer, primary purposes of use such as home safety and control, and attributes not associated with the person, contribute heavily toward an "Allow" decision.

4.2.2 Situational Factors

Model Selection and Evaluation. Because the goal of this model is to identify situational factors that could make users more or less comfortable with a given information flow for which a preference was already identified, we used the same features from the best model in the allow/deny prediction problem, but also added the situational factors as features. The validation performance (AUC) with this model was .895 and .912 with PySpark and scikit-learn, respectively. 10-fold cross-validation for this model in scikit-learn resulted in a .907 AUC. The selected model had a test set AUC of .941 and .899 with PySpark and scikit-learn, respectively.

Predicting Changes To demonstrate how our model can be used in a combinatorial way, we first took the average of the numerical features (e.g., average comfort manufacturer, average IUIPC values) to create a hypothetical "average user" from our data set, and predicted the comfort change from information flows with "13 choose k" situational factors that could change this user's comfort level toward different information flows. Because our survey allowed users to select at most three situational factors out of 13, we set $k = 3$. Then we created scenarios with the Cartesian product of the numerical features (e.g., IUIPC constructs, mean comfort toward manufacturer, etc) $\times$ attributes $\times$ purposes $\times$ devices $\times$ 13 choose 3 situational factors, resulting in 484,484 scenarios for this hypothetical user. In doing so, we learned to what extent certain situational factors, when selected, "matter" to each specific scenario for this user, as well as in general, what situational factors would matter for this user, considering all combinations of attributes, purposes, and devices. The extent and direction to which they "matter" is expressed by the percentage of information flows involving specified attributes, purposes, and situational factors, which would make users more or less comfortable. In other words, the percentage of predictions in each class will indicate whether the presence of a certain situational factor will make a user more or less comfortable. For example, if *"used for primary purposes or not?"* is checked and the percentage of predictions for "less comfortable" is greater than the percentage of predictions for "more comfortable" then it means that for this user, secondary uses

#	Attribute	Purpose	What if... (situational factor selected)	% more comfortable	% less comfortable
1	Any	Any	user can control or not?	**84.6%**	15.4%
2	Any	Any	data handled securely or not?	43.9%	**56.1%**
3	Any	Any	used only for primary purposes or not?	32.9%	**67.1%**
4	Any	Any	user is aware or not?	**63.3%**	36.7%
5	Any	Any	used for safety or not?	**69.8%**	30.2%
6	Any	Targeted ads	user can control or not?	**95.5%**	4.5%
7	Any	Targeted ads	user is aware or not?	**79.9%**	20.1%
8	Indoor location	Any	used only for primary purposes or not?	36.2%	**63.8%**
9	Indoor location	Any	user can control or not?	**92.1%**	7.9%
10	Indoor location	Any	user has consented or not?	**57.1%**	42.9%
11	Indoor location	Any	manufacturer well known or not?	**51.8%**	48.2%
12	Age of people at home	Any	manufacturer well known or not?	**73.6%**	26.4%
13	Any	Home safety	manufacturer well known or not?	13.8%	**86.2%**
14	Any	Home safety	used only for primary purposes or not?	5.5%	**94.5%**
15	Energy use	Targeted ads	user is aware or not?	**68%**	32%
16	Energy use	Targeted ads	user can benefit or not?	**86.8%**	13.2%

Table 2. Model predictions for the "average" user in the scenarios data set. Percentages indicate the number of scenarios for which "more" or "less" comfortable was predicted when the situational factor is present i.e. "checked." Comfort changes can be identified using combinations within information flows considering specific attributes, purposes, or devices, as well as for any levels of such factors.

may make them less comfortable. In another example, if *"user can control or not?"* resulted in a greater number of "more comfortable" predictions, it means that for this situational factor, this user would be more comfortable if they could control the information flow.

Identifying Actionable Steps. Table 2 contains some examples, indicating the percentage of scenarios for which the prediction was "more comfortable" and otherwise (i.e., "less comfortable"). By using the model in this way, a developer can learn that, for example, it would make the user more comfortable in general if they could control the information flows (84.6%, #1). Also from the table, one can see that if data are used beyond primary purposes, the user would be uncomfortable (67.1%, #3). Another example, involving specific attributes, can be extracted from comparing the difference between two situational factors: for information flows involving energy use for targeted ads, a higher number of scenarios were predicted as "more comfortable" for "user can benefit," (86.8%, #16) than for "user is aware," (68%, #15), so for this user to be more comfortable, benefit seems more important than awareness.

This experiment demonstrates how not only overall comfort changes can be predicted from using the situational factors, but also how relevant, fine-grained changes can be identified by considering the predictions within an information flow involving a particular attribute, purpose, or device. These predictions can inform developers toward actionable steps in preserving the privacy of their users and developing user trust. For instance, from Table 2, some actions would be that, if the indoor location is to be used by the developer for any purpose, then the manufacturer should first and foremost give user control (92.1% more comfortable, #9), not use it for secondary purposes (63.8% less comfortable, #8), and get user consent (57.1% more comfortable, #10). It would also be OK for the manufacturer to use any data for the safety of the home (69.8%, #5).

4.2.3 Cost of Privacy

Model Selection and Evaluation. When selecting the best model for predicting the dollar amount under the four circumstances, namely pay extra when purchasing, discount at purchase, refund after purchase, and one-time fee after purchase, the best-performing model with both libraries was a linear regression (implemented with stochastic gradient descent in scikit-learn) using the following features: IUIPC scores, economic scenario (one of the four), and average comfort level toward manufacturer (scikit-learn only). The RMSE for the validation set was 15.772 and 14.666 for PySpark and scikit-learn, respectively. 10-fold cross validation with scikit-learn resulted in a RMSE of 15.101. The performance of the selected model on the test set (RMSE) was 14.292 and 12.459 for PySpark and scikit-learn, respectively. Figure 5 shows the predictions made on the test set.

Predicting Privacy Value. Our model could be used by collecting the IUIPC scores, average comfort

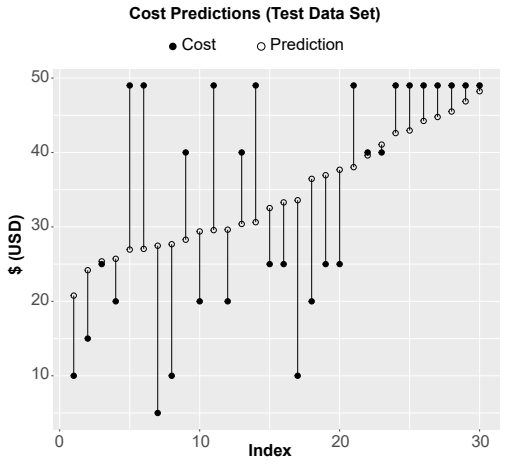

Fig. 5. Cost predictions on the test set using the best model considering a $49 device. Points are ordered by the prediction value.

levels, average notification frequency, and automatically feeding the four circumstances and predicting the dollar amount for each economic scenario in order to understand how much users would be willing to pay for extra privacy protections and/or willing to take as a discount/refund in order to allow more data collection and sharing. For example, using the hypothetical "average user" from our data set, when purchasing a voice assistant costing $49, this user would take a discount of $38.03 in exchange for sharing their data, but would only pay $31.24 to protect such data. Similarly, the predictions indicate that after purchasing the device, this user would take a refund of $44.48 in exchange for fewer privacy protections, but would only be willing to pay $28.01 as a fee to protect their privacy after having purchased the device. Translating this into practical setting, a developer could estimate the dollar value of new, optional privacy features before introducing them.

5 Discussion

Our findings suggest that (1) attributes associated with the person (not the home) and that hint at habits and lifestyle cause user discomfort; (2) appropriation and secondary use may not be acceptable in the smart home; (3) potential smart home users would assign a lower monetary value to privacy when they do not have it than when they do; (4) most users would not be willing to take discounts/refunds to give away privacy and would expect privacy not to cost extra money and (5) personalized privacy preferences about information flows and their changes can be predicted with machine learning.

Secondary Uses. As seen through our work, it may not be OK for data practices from other domains such as the Web and mobile apps to transfer into the smart home. The smart home differs largely from those domains, with a major distinction being that in most cases, popular smart home products are paid for, whereas the most popular websites and smartphone apps are used for free, and the reason why they are sustainable is because of practices such as online behavioral advertising and the sharing of personal data with third party organizations. Nonetheless, major companies such as Google and Amazon are behind the most popular smart home devices such as voice assistants, thermostats, and smart locks, which creates opportunities for data collected through these devices to be part of the larger ecosystem composed by technologies from other domains in which these companies do business in (e.g., targeted advertising, online shopping). In light of our findings, this practice would be unacceptable because users do not want their data to be collected or shared for non-primary purposes, given that they already paid for a smart home device.

Privacy Value. If secondary uses such as targeted advertising are deemed inappropriate, and adoption of devices sold by leaders in such practice is growing, then the privacy paradox already applies. For instance, while prior work and our own study suggest that users highly value privacy and would be concerned about privacy when considering smart home devices, adoption of such devices has significantly increased in recent years – even in our study, nearly half of the participants reported already having a smart home device, with 29% (201) indicating that they own a voice assistant. This suggests that optimism bias [3] and the privacy paradox may facilitate the establishment of practices that users do not deem acceptable, yet consumers will either tolerate them in exchange for convenience or will be unaware of them. What's more concerning about this landscape is the power imbalance and information asymmetry [16] that may quietly arise, making the smart home an environment where privacy is not included by default, which is aggravated by what our findings suggest, in line with works in other domains: people would value privacy less when they do not have it. If consumers are not willing to pay extra for privacy, then the economic incentive for manufacturers to include these protections is somewhat limited. Nonetheless, our findings suggest that many will either expect privacy protections to be included by default or would not take any money to give it away in the case where such protections are already in place.

Contextual Integrity. Given these circumstances, then how designers, developers, and manufacturers could respect established social boundaries of the home [30, 36] while being able to offer products and services that make their customers' lives easier without infringing on their privacy expectations? This is an ethical question because what is known so far seems to suggest that developers of smart home products will be able to penetrate markets regardless of how they use their customers' data, practicing business models originating in other domains such as the Web, when users prioritize convenience and underestimate risks. However, a large number of consumers are worried about privacy, and as our findings indicate, they would expect privacy protections to be embedded into the devices they purchase, and furthermore, would be willing to assign a relatively high monetary value to protect their personal data. To serve such users, manufacturers could incorporate and market privacy by design, which could lead to acceptable transactions where both sides benefit without posing major risks to individuals and society at large. This is precisely where our modeling work fits in, so that consumer privacy preferences are in line with business practices of smart home developers. This creates a design space for consumer-facing (e.g., [6, 12]) and developer-facing tools (e.g. our own).

Modeling Personalized Preferences. By replicating our work, developers could model privacy preferences based on measurements of concern toward internet privacy such as the IUIPC, in addition to comfort levels for four randomly generated scenarios. For instance, while our survey took a median time of 19 minutes to complete, including explanations and demographics, a user might be able to provide the data needed in much less time. Making this a quick activity is important, since answering such questions may not be the primary goal of users, so they are likely to skip them if they take too long or are too complicated. Such data could be collected when a user first installs a smart home device in order to inform developers about what may be appropriate/inappropriate for a given user. We also envision a scenario where such models could be used in the process of purchasing a smart home device that would be aligned with user privacy preferences.

Predicting Changes. Beyond modeling initial allow/deny preferences, we also show that a model could be used to identify which situations could change a user's comfort upward or downward given different information flows. This can give developers actionable steps toward making users more comfortable as well as understanding what situations should be avoided that would cause privacy concerns and distrust. For example, manufacturers could emphasize that certain data are being collected only for primary purposes, or collect data much less frequently. Doing so could lead to more awareness and control, mitigation of privacy concerns, and increased user trust toward the device/developer.

Employing Privacy Values. If developers truly need to leverage user data for business purposes, it would be possible to know how much a user would be willing to pay for extra privacy protections, as well as how much they would be willing to accept in exchange for being more liberal about data collection and sharing in the smart home. While limited to our own scenario with a voice assistant costing $49, we show through our methodology that it is possible to predict the "cost of privacy" in smart home scenarios. Our value prediction model could also be leveraged by retailers when helping users choose the right devices, by looking at the amounts derived by the model and choosing a device that reflects their privacy concerns and valuations. We note, however, that our approach to predict costs was not as effective as the approach to predict preferences, which may indicate that the features that worked for modeling allow/deny and predicting situational factors did not work as well for modeling privacy values. We encourage further exploration on this topic.

Limitations. We acknowledge the limitations of working with users from AMT (also known as Turkers) for our online survey. For example, while there's diversity of age, gender, and income, the sample is not representative of the US population and there could be limitations about skewed education levels [35]. We also acknowledge that Turkers may skew toward people with non-traditional employment or underemployment who may stay at home more often and/or have more experience with information technology and the Internet.

We note that because cultural differences are known to affect privacy preferences, the generalization of our models are sensitive to the context in which the training data was obtained, which in our case, involves people in the United States who participate on Amazon Mechanical Turk (AMT) tasks, who may or may not have smart home devices. Nonetheless, we provide the demographics of survey participants, which indicates that nearly half of them already have a smart home device.

We also acknowledge that our survey measured subjective comfort levels and hypothetical scenarios, which we use as a proxy for how concerned users may be in regards to data collection practices in the smart home, as well as to identify acceptable practices. In the field of privacy research, people's attitudes and behaviors are

known to differ, which reflects the privacy paradox [23]. Because of this, we interpreted our results taking this into consideration, and hope that our findings can inform and educate manufacturers, policy makers, and end-users regarding the future of smart homes. Nonetheless, the use of data representing attitudes is acceptable in investigating privacy expectations and concerns in reasonably new domains in a scalable manner [5].

In predicting changes to preferences, our approach only predicts the direction of changes and the relative contribution of a situational factor toward such changes. It would be useful to also predict what could make people accept or deny an information flow, in other words, what could change their mind.

6 Conclusion

Public discourse and consumer concerns around the privacy of smart home devices are commonplace because they challenge a long-held notion of the home as a private and intimate place. Therefore, identifying appropriate data practices is an important step toward safeguarding the home's privacy and developing user trust.

We present the design and evaluation of machine learning models to derive privacy preferences and changes to such preferences in the smart home, considering the many contextual factors known to influence privacy decisions. We show that through a short survey obtaining responses to four random information flows and the IUIPC scale, our model can predict allow/deny preferences (AUC .868), along with situations that could make users more or less comfortable (AUC .899). We also describe our attempt to create a model to predict the dollar amount users would pay/accept in exchange for privacy in the smart home (RMSE 12.459).

Our work enables smart home developers to preserve the privacy of their users and take steps toward building user trust in the smart home.

7 Acknowledgements

We thank our survey participants for their valuable input. We also thank the anonymous reviewers for their insightful comments and suggestions. We thank our shepherd, Jens Grossklags, for providing useful guidance and feedback in the revision process.

We acknowledge Daniel Acuña for his invaluable guidance on the development of our machine learning models, and the people of the SALT Lab in the School of Information Studies at Syracuse University. This work was supported in part by the National Science Foundation (NSF) grant number CNS-1464347.

References

[1] Alessandro Acquisti. 2002. Protecting privacy with economics: Economic incentives for preventive technologies in ubiquitous computing environments. In *Proceedings of Workshop on Socially-informed Design of Privacy-enhancing Solutions, UbiComp 2002*.

[2] Alessandro Acquisti, Idris Adjerid, Rebecca Balebako, Laura Brandimarte, Lorrie Faith Cranor, Saranga Komanduri, Pedro Giovanni Leon, Norman Sadeh, Florian Schaub, Manya Sleeper, et al. 2017. Nudges for privacy and security: Understanding and assisting users' choices online. *CSUR* 50, 3 (2017), 44.

[3] Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. 2015. Privacy and human behavior in the age of information. *Science* 347, 6221 (2015), 509–514.

[4] Alessandro Acquisti, Leslie K John, and George Loewenstein. 2013. What is privacy worth? *The Journal of Legal Studies* 42, 2 (2013), 249–274.

[5] Noah Apthorpe, Yan Shvartzshnaider, Arunesh Mathur, Dillon Reisman, and Nick Feamster. 2018. Discovering smart home internet of things privacy norms using contextual integrity. *Proceedings of IMWUT* 2, 2 (2018), 59.

[6] Paritosh Bahirat, Yangyang He, Abhilash Menon, and Bart Knijnenburg. 2018. A data-driven approach to developing IoT privacy-setting interfaces. In *IUI 2018*. ACM, 165–176.

[7] Eun Kyoung Choe, Sunny Consolvo, Jaeyeon Jung, Beverly Harrison, and Julie A Kientz. 2011. Living in a glass house: a survey of private moments in the home. In *Proceedings of UbiComp 2011*. ACM, 41–44.

[8] CIPR 2017. CIPR - Home automation device market grows briskly. (2017). https://www.voicebot.ai/wp-content/uploads/2017/11/cirp-news-release-2017-11-06-echo-home.pdf.

[9] Bogdan Copos, Karl Levitt, Matt Bishop, and Jeff Rowe. 2016. Is anybody home? Inferring activity from smart home network traffic. In *SPW, 2016*. IEEE, 245–251.

[10] Karen L Courtney. 2008. Privacy and senior willingness to adopt smart home information technology in residential care facilities. *Methods of Information in Medicine* 47, 01 (2008), 76–81.

[11] Karen L Courtney, George Demeris, Marilyn Rantz, and Marjorie Skubic. 2008. Needing smart home technologies: the perspectives of older adults in continuing care retirement communities. (2008).

[12] Anupam Das, Martin Degeling, Daniel Smullen, and Norman Sadeh. 2018. Personalized privacy assistants for the internet of things: providing users with notice and choice. *IEEE Pervasive Computing* 17, 3 (2018), 35–46.

[13] George Demiris, Brian K Hensel, Marjorie Skubic, and Marilyn Rantz. 2008. Senior residents' perceived need of and preferences for "smart home" sensor technologies. *Interna-*

tional Journal of Technology Assessment in Health Care 24, 1 (2008), 120–124.

[14] Jens Grossklags and Alessandro Acquisti. 2007. When 25 cents is too much: An experiment on willingness-to-sell and willingness-to-protect personal information. In *WEIS 2007*.

[15] Jason Hong. 2017. The privacy landscape of pervasive computing. *IEEE Pervasive Computing* 16, 3 (2017), 40–48.

[16] Xiaodong Jiang, Jason I Hong, and James A Landay. 2002. Approximate information flows: Socially-based modeling of privacy in ubiquitous computing. In *UbiComp 2002*. Springer, 176–193.

[17] Juniper 2017. Juniper - digital voice assistants. (2017). https://www.juniperresearch.com/researchstore/innovation-disruption/digital-voice-assistants/platforms-revenues-opportunities.

[18] Bart P Knijnenburg, Alfred Kobsa, and Hongxia Jin. 2013. Dimensionality of information disclosure behavior. *International Journal of Human-Computer Studies* 71, 12 (2013), 1144–1162.

[19] Scott Lederer, Jennifer Mankoff, and Anind K Dey. 2003. Who wants to know what when? privacy preference determinants in ubiquitous computing. In *CHI'03 extended abstracts on Human factors in computing systems*. ACM, 724–725.

[20] Hosub Lee and Alfred Kobsa. 2016. Understanding user privacy in Internet of Things environments. In *IEEE WF-IoT 2016*. 407–412.

[21] 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 UbiComp 2012*. ACM, 501–510.

[22] 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 privacy assistant for mobile app permissions. In *SOUPS 2016*. 27–41.

[23] Bernard Lubin and Roger L Harrison. 1964. Predicting small group behavior with the self-disclosure inventory. *Psychological Reports* 15, 1 (1964), 77–78.

[24] 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.

[25] Emily McReynolds, Sarah Hubbard, Timothy Lau, Aditya Saraf, Maya Cakmak, and Franziska Roesner. 2017. Toys that listen: A study of parents, children, and internet-connected toys. In *Proceedings of CHI 2017*. 5197–5207.

[26] William Melicher, Mahmood Sharif, Joshua Tan, Lujo Bauer, Mihai Christodorescu, and Pedro Giovanni Leon. 2016. (Do Not) Track me sometimes: users' contextual preferences for web tracking. *PETS* 2016, 2 (2016), 135–154.

[27] Pardis Emami Naeini, Sruti Bhagavatula, Hana Habib, Martin Degeling, Lujo Bauer, Lorrie Cranor, and Norman Sadeh. 2017. Privacy expectations and preferences in an IoT world. In *SOUPS 2017*.

[28] Helen Nissenbaum. 2004. Privacy as contextual integrity. *Wash. L. Rev.* 79 (2004), 119.

[29] Katarzyna Olejnik, Italo Dacosta, Joana Soares Machado, Kévin Huguenin, Mohammad Emtiyaz Khan, and Jean-Pierre Hubaux. 2017. SmarPer: Context-aware and automatic runtime-permissions for mobile devices. In *IEEE SP 2017*. 1058–1076.

[30] Leysia Palen and Paul Dourish. 2003. Unpacking privacy for a networked world. In *CHI 2003*. ACM, 129–136.

[31] Pew 2 2017. Pew Research Center, The internet of things connectivity binge: what are the implications? (2017). http://www.pewinternet.org/2017/06/06/theme-3-risk-is-part-of-life-the-internet-of-things-will-be-accepted-despite-dangers-because-most-people-believe-the-worst-case-scenario-would-never-happen-to-them.

[32] Yu Pu and Jens Grossklags. 2015. Using conjoint analysis to investigate the value of interdependent privacy in social app adoption scenarios. (2015).

[33] Yu Pu and Jens Grossklags. 2016. Towards a model on the factors influencing social app users' valuation of interdependent privacy. *PETS* 2016, 2 (2016), 61–81.

[34] Yu Pu and Jens Grossklags. 2017. Valuating friends' privacy: Does anonymity of sharing personal data matter?. In *SOUPS 2017*. 339–355.

[35] Joel Ross, Lilly Irani, M Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. Who are the crowdworkers? shifting demographics in mechanical turk. In *CHI'10 extended abstracts on Human factors in computing systems*. ACM, 2863–2872.

[36] Daniel J Solove. 2005. A taxonomy of privacy. *U. Pa. L. Rev.* 154 (2005), 477.

[37] Blase Ur, Pedro Giovanni Leon, Lorrie Faith Cranor, Richard Shay, and Yang Wang. 2012. Smart, useful, scary, creepy: perceptions of online behavioral advertising. In *Proceedings of SOUPS 2012*. ACM, 4.

[38] Max Van Kleek, Reuben Binns, Jun Zhao, Adam Slack, Sauyon Lee, Dean Ottewell, and Nigel Shadbolt. 2018. X-ray refine: Supporting the exploration and refinement of information exposure resulting from smartphone apps. In *Proceedings of CHI 2018*. ACM, 393.

[39] Primal Wijesekera, Arjun Baokar, Lynn Tsai, Joel Reardon, Serge Egelman, David Wagner, and Konstantin Beznosov. 2017. The feasibility of dynamically granted permissions: Aligning mobile privacy with user preferences. In *IEEE SP 2017*. 1077–1093.

[40] Primal Wijesekera, Joel Reardon, Irwin Reyes, Lynn Tsai, Jung-Wei Chen, Nathan Good, David Wagner, Konstantin Beznosov, and Serge Egelman. 2018. Contextualizing privacy decisions for better prediction (and protection). In *Proceedings of CHI 2018*. ACM, 268.

[41] Peter Worthy, Ben Matthews, and Stephen Viller. 2016. Trust me: doubts and concerns living with the internet of things. In *Proceedings of DIS 2016*. ACM, 427–434.

A Appendix

A.1 Survey Details

A.1.1 Scenario Components

The vignettes used in our survey were randomly generated from a list of contextual factors, namely attributes, purposes, and devices. Random generation of information flows is a practice relied upon by prior works investigating IoT preferences [5, 20, 27], and allows the collection of responses for a large combination of factors. While we acknowledge that some combinations may not make sense at first, such as age of people at home from doorbell camera for targeted advertising, learning such preferences can protect the user from data practices that may attempt to use the data in unexpected or unforeseen ways in the future. In the review step, situational factors were used to understand which ones could make users more or less comfortable with the given information flow. Below we explain our component choices.

Devices. We identified the most popular smart home devices on shopping websites such as Amazon and Best Buy, as well as popular consumer blogs. From this step, we identified a list of six devices: doorbell camera, smart lights, smart lock, security camera, smart thermostat, and voice assistant/hub.

Attributes. We looked at product descriptions from devices identified in the previous step in order to understand their advertised features and identify potential personal and home attributes that could be directly collected as well as those that could be inferred, since inferred attributes were deemed more concerning by users in the context of the IoT [27]. For example, we posited that the developer/manufacturer of a doorbell camera could infer the number of people at home at a certain point in time. We also posited that the manufacturer of a security camera could infer the gender and age of people inside the home. Likewise, a smart lock could hint at habits and lifestyle, such as the time a person leaves and returns to their home. Other attributes were determined directly from the description of such products, for example, *"voice-control your music"* is a feature described for Amazon's Alexa, resulting in the "Music, shows, or movies" attribute. Other attributes could be determined indirectly, such as one's home location being obtained from the device's public IP address. Table 3 (appendix) shows all attributes used in our survey along with their descriptions as shown in the survey.

Purposes. For purposes of data use, we considered those that are primary to the smart home, such as home control, home automation, and home safety/security, in addition to purposes identified in prior work on online behavioral advertising [26], based on the rationale that the smart home can be seen as an extension of the web, since devices are connected and controlled over the Internet, and manufacturers of the most popular smart home devices are major Internet companies (e.g., Google and Amazon). Table 3 (appendix) shows all the purposes used in our survey along with their respective descriptions, exactly as seen by participants. While some purposes are closely related, for example, identity linking and targeted advertising, each purpose of use was presented alongside their description to prevent any ambiguity and overlap. For example, for identity linking, the description is focused around the association of data with one's identity, and for targeted advertising, the description focuses on tailored offering of products.

Situational Factors. Finally, we also considered situational factors that are known to affect users' privacy preferences in other application domains [21, 26, 37], such as mobile apps and Web browsing. These situational factors were presented in our survey in order to identify what factors could sway participants from their original preferences. For example, when they provided their subjective comfort level with a scenario, we asked them to select up to three circumstances that could change their preferences, making them either more or less comfortable. For example, if they indicated being comfortable with a certain scenario (3-5 score), we would ask them what circumstances would make them *less* comfortable, such as *"if data were not handled securely," "if data were used beyond primary purposes,"* or *"if the manufacturer/developer was unknown."* Table 3 (appendix) shows all the situational factors used in our survey, along with their respective descriptions.

A.2 Machine Learning Details

Data Preparation. We made modifications to our data sets in order to prepare them for our machine learning pipeline. For the first data set, we added the average comfort level grouped by participant to each row, for manufacturer, third party, government, and identity. We also added to each row the average notification level given by each participant (e.g., avg_comfort_manufacturer, avg_comfort_third_party, avg_notified, etc.). We did this because in a practical setting, we ideally would like

Attribute	Description
Activity	what you do inside your home such as cooking, studying, singing, exercising
Age of people at home	the age of all the people who visit and live in your home
Apps used	apps that you downloaded to perform functions on your TV, voice assistant, or mobile phone
Calendar events, alarms, and timers	calendar events or reminders that you have set up on voice assistants
Communications	calls made with your smart devices or text messages you sent or received from others
Destinations	places you visit immediately after leaving your home
Device actions	when the device is switched on or off, when the device is used or controlled
Device brand/model	the manufacturer, model, and make of your device
Device events	when the device's sensors are activated due to activity in the home
Device states	the current status of your device such as whether it is on or off, activated or deactivated, lock or unlocked, open or closed
Energy use	how much energy you are currently using as well as your energy use history
Gender of people at home	the gender of all the people who visit or live in your home
Habits and Lifestyle	how frequently you shop, eat out, travel, and do other things indicative of your lifestyle
Indoor location	the precise location such as the room you are in (e.g., bathroom, living room)
Inside temperature	the temperature inside of your home
Music, shows, or movies	entertainment that you may engage with through smart speakers, voice assistants, smart TVs, and gaming consoles
Noise levels	the level of auditory noise and activity inside your home
Number of people at home	the number of people that live in and visit your home
Outside temperature	the temperature outside of your home
Sleep data	the number of hours slept and the quality of your sleep, including history data
Weather	outside climate features such as whether it is cloudy, rainy, snowy, etc.

Purpose	Description
Company revenue	for the profit of a company who is behind your smart device (e.g., manufacturer, retailer, etc.)
Customized experiences/personalization	to save you time and recommend/target content and features based on your needs
Home automation	to automate how items in and around the house work without your intervention
Home control	to switch devices on and off or manage and control objects, appliances, and electronics in your home
Home safety/security	to ensure the safety and physical security of your home or in case of an emergency
Identity linking	to associate other collected data with your identity
Legal actions	to use your data for a lawsuit that you may or may not be involved in
Price discrimination	to give you discounts, sales, coupons, or determine the price of something based on your needs
Targeted ads	to suggest products and services most tailored to you
User tracking and profiling	to create a virtual profile of your person that most accurately represents you

Situational Factor	More comfortable	Less comfortable
Entity	If the manufacturer was well known	If the manufacturer was unknown
Consent	If I gave consent to collect data	If I did not give consent to collect data
Frequency	If information was collected less frequently	If information was collected more frequently
Sensitive	If the information involved was not sensitive	If the information involved was sensitive
Benefit	If I could benefit from it (e.g., discounts, serendipitous opportunities)	If I could not benefit from it (e.g., discounts, serendipitous opportunities)
Retention	If the information was stored for a short period of time, then deleted	If the information was stored for a longer period of time, or never deleted
Purpose	If the information was only used for the intended purpose	If the information was used beyond the intended purpose
Awareness	If I was aware of how the data were being used	If I was not aware of how the data were being used
Safety	If the data collection was useful for personal and home safety	If the data collection was not useful for personal and home safety
Improvement	If the data were used for improving products and services	If the data were not used for improving products and services
Common Good	If the data were used for the common good (e.g., benefit the society at large)	If the data were not used for the common good (e.g., benefit the society at large)
Control	If I could control the data (e.g., access, copy, and delete)	If I could not control the data (e.g., access, copy, and delete)
Secure	If data were handled and secured properly	If data were not handled and secured properly

Table 3. Attributes, Purposes, and Situational Factors used in survey, along with examples provided exactly as seen by participants.

Survey Participant Demographics			
Gender		**Own SH device**	
Female	339 (48.6%)	No	356 (51%)
Male	356 (51%)	Yes	342 (49%)
Other	3 (0.4%)		
Age		**Education**	
18-25	113 (16.2%)	< High school	1 (0.1%)
26-35	315 (45.1%)	High school	60 (8.6%)
36-45	147 (21.1%)	Associate	89 (12.8%)
46-55	69 (9.9%)	Some college	141 (20.2%)
56-65	54 (7.7%)	Bachelor's	285 (40.8%)
>65	9 (1.3%)	Professional	18 (2.6%)
		Master's	97 (13.9%)
		Doctoral	7 (1%)
Income		**IUIPC Scale**	
<10k	31 (4.4%)	*Control*	
10k-39k	211 (30.2%)	Mean [SD]	6.01 [0.99]
40k-69k	210 (30.1%)	Median	6
70k-100k	134 (19.2%)	*Awareness*	
100k-149k	74 (10.6%)	Mean [SD]	6.38 [0.94]
>150k	38 (5.4%)	Median	7
		Collection	
Have children		Mean [SD]	5.9 [1.15]
No	381 (54.6%)	Median	6.25
Yes	317 (45.4%)		
Weekly Internet use (hours)		**Marital Status**	
Mean [SD]	43.41 [28.64]	Never married	307 (44%)
Median	39	Married	325 (46.6%)
		Separated	10 (1.4%)
		Divorced	51 (7.3%)
		Windowed	5 (0.7%)

Table 4. Demographics of survey participants. 49% of participants claimed to own a smart home device. SH = Smart Home.

PySpark's MLlib			
Top Coefficients Deny		**Top Coefficients Allow**	
Legal Actions(P)	-1.026	Comfort Manuf.	1.237
Communications(A)	-.960	Outside Temp.(A)	.787
Identity Linking(P)	-.952	Weather(A)	.737
Age of people(A)	-.798	Inside Temp.(A)	.668
Targeted Ads(P)	-.727	Personalization(P)	.632
Gender of people(A)	-.662	Home Safety(P)	.543
Destinations(A)	-.533	Device model(A)	.528
Tracking and Profiling(P)	-.442	Energy use(A)	.414
Company Revenue(P)	-.436	Smart Lights(D)	.381
Noise levels(A)	-.239	Home control(P)	.338
scikit-learn			
Top Coefficients Deny		**Top Coefficients Allow**	
Communications(A)	-1.251	Comfort Manuf.	6.346
Legal Actions(P)	-1.139	Inside Temp.(A)	1.191
Age of people(A)	-1.128	Weather(A)	1.030
Identity Linking(P)	-.828	Home Safety(P)	.943
Targeted Ads(P)	-.760	Outside Temp.(A)	.932
Gender of people(A)	-.718	Personalization(P)	.814
Destinations(A)	-.634	Home Control(P)	.742
IUIPC Collection	-.565	Energy Use(A)	.639
Habits and Lifestyle(A)	-.380	Not specified(P)	.625
Not specified(D)	-.320	IUIPC Awareness	.527

Table 5. Top coefficients toward either "Deny" (negative) or "Allow" (positive). A = attribute, P = purpose, D = device.

to calculate comfort levels from a small number of scenarios, and be able to use it to predict preferences for a large number of scenarios (e.g., number of purposes X number of attributes X number of devices, given average comfort levels from four scenarios). Considering prior works by Bahirat *et al.* [6], we also modified this data set to include the cluster value of participants. We included columns from 5-to-3 clusters derived from k-means clustering from IUIPC constructs alone, then adding entity comfort levels (i.e., manufacturer, third party, and government), adding identity comfort, and adding notification level as features for the clustering. Given their results, we wanted to compare if a clustered approach would yield better results. For the second data set (economics data set), we added the IUIPC values and the average comfort levels from the participant for manufacturer, third party, government, and identity, as well as the notification frequency. Given the loss aversion observed in previous works in regards to privacy valuations (e.g., [14]), we also added a categorical variable indicating whether the user already "had" privacy or not, according to the scenario condition. Finally, we removed any rows indicating a dollar amount greater than the price of the voice assistant, that is, greater than \$49 or equal to \$0 (273 rows), because they would prevent the model from capturing realistic discounts/costs.

Feature Engineering. For both data sets, numerical features (e.g., IUIPC) were scaled with min-max scaling[2]. For categorical features such as attributes, purposes of data use, and devices, the columns were one-hot encoded[3], with missing values representing a category.

2 https://en.wikipedia.org/wiki/Feature_scaling#Rescaling_(min-max_normalization)

3 https://en.wikipedia.org/wiki/One-hot

Predict Allow/Deny				
Features	**Best Algorithm**		**AUC**	
	PySpark MLlib	**scikit-learn**	**PySpark MLlib**	**scikit-learn**
IUIPC	Random Forest	Logistic Regression	0.656	0.646
+ Attribute	Random Forest	Logistic Regression	0.685	0.674
+ Purpose	Logistic Regression	Logistic Regression	0.723	0.736
+ Device	Logistic Regression	Logistic Regression	0.731	0.739
IUIPC	Random Forest	Logistic Regression	0.656	0.646
+ Comfort toward Manufacturer	Logistic Regression	Multilayer Perceptron	0.829	0.817
+ Comfort toward Third Party	Logistic Regression	Support Vector Machine	0.827	0.816
+ Comfort toward Government	Logistic Regression	Stochastic Gradient Descent	0.828	0.816
+ Comfort with Identity	Logistic Regression	Stochastic Gradient Descent	0.825	0.814
+ Notification Frequency	Logistic Regression	Multilayer Perceptron	0.826	0.817
IUIPC, Attribute, Purpose, Device	Logistic Regression	Logistic Regression	0.731	0.739
+ Comfort toward Manufacturer	Logistic Regression	Logistic Regression	0.86	**0.861**
+ Comfort toward Third Party	Logistic Regression	Logistic Regression	0.86	0.861
+ Comfort toward Government	Logistic Regression	Logistic Regression	**0.861**	0.861
+ Comfort with Identity	Logistic Regression	Logistic Regression	0.858	0.859
+ Notification Frequency	Logistic Regression	Logistic Regression	0.858	0.859
Predict Preference Changes				
Features	**Best Algorithm**		**AUC**	
	PySpark MLlib	**scikit-learn**	**PySpark MLlib**	**scikit-learn**
IUIPC, Attribute, Purpose, Device	-	-	-	-
+ Comfort toward Manufacturer	-	Logistic Regression	-	**0.912**
+ Comfort toward Third Party	-	-	-	-
+ Comfort toward Government	Logistic Regression	-	**0.895**	-
Predict Privacy Value				
Features	**Best Algorithm**		**RMSE**	
	PySpark MLlib	**scikit-learn**	**PySpark MLlib**	**scikit-learn**
IUIPC	Linear Regression	Stochastic Gradient Descent	16.541	15.814
+ Economic Scenario	Linear Regression	Stochastic Gradient Descent	**15.683**	14.803
+ Comfort toward Manufacturer	Linear Regression	Stochastic Gradient Descent	15.772	**14.666**
+ Comfort toward Third Party	Linear Regression	Stochastic Gradient Descent	15.964	14.769
+ Comfort toward Government	Linear Regression	Stochastic Gradient Descent	15.844	14.733
+ Comfort with Identity	Linear Regression	Stochastic Gradient Descent	15.974	15.07
+ Notification Frequency	Linear Regression	Stochastic Gradient Descent	16.253	15.121

Table 6. Results from model selection based on best performance on validation set. Same features from Allow/Deny were used to predict comfort change for each scenario. Bold text indicates best value.