ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230614

Privacy-Enhanced AI Assistants Based on Dialogues and Case Similarity

Xiao Zhan[a;*], **Ștefan Sarkadi**[a] **and Jose Such**[a,b]

[a]King's College London, UK
[b]VRAIN, Universitat Politecnica de Valencia, Spain

Abstract. Personal assistants (PAs) such as Amazon Alexa, Google Assistant and Apple Siri are now widespread. However, without adequate safeguards and controls their use may lead to privacy risks and violations. In this paper, we propose a model for privacy-enhancing PAs. The model is an *interpretable* AI architecture that combines 1) a dialogue mechanism for understanding the user and getting online feedback from them, with 2) a decision-making mechanism based on case-based reasoning considering both user and scenario similarity. We evaluate our model using real data about users' privacy preferences, and compare its accuracy and demand for user involvement with both online machine learning and other, more interpretable, AI approaches. Our results show that our proposed architecture is more accurate and requires less intervention from the users than existing approaches.

1 Introduction

AI assistants such as Personal Assistants (PAs) have become a key application of AI techniques. Over the last decade, they have become widespread in our homes and our phones, including Amazon Alexa, Google Assistant, Apple Siri, and so on. Despite their popularity and the convenience and functionalities they offer to users, PAs have also raised significant concerns regarding end users' privacy [2, 16, 3]. PAs have a distinct working ecosystem of their own, which is complicated and involves many different stakeholders [6, 2, 3]. For instance, PAs depend on cloud service providers to store their data. Additionally, to provide their vast range of services, they use both built-in skills and third-party applications called skills [14, 5]. The disadvantage of this complex ecosystem is that users' personal information may be accessed or misused by unauthorised parties without the user's awareness [44, 15, 10]. For instance, this may occur when Spotify's service provider accesses users' login details while playing music via Alexa [6], and everyone within audible range may know the status of a smart door lock [34].

Most PA users have inaccurate mental models of the interactions between the different stakeholders in a PA's ecosystem and lack adequate mechanisms to take control of their privacy [2]. At the same time, when those interactions are made apparent to users and promising privacy protection mechanisms suggested in previous studies are given to them, such as access control mechanisms [50], those mechanisms end up not being utilized in practice because users find it too burdensome [50]. In particular, although all users in a previous study wanted to have protection mechanisms and wanted to exert control over the flows of information, they did not want to spend the time setting the mechanisms up because it was considered inconvenient [50]. Instead, they expected the PA to quickly learn what the social norms regarding privacy were while intervening the least possible. This is in line with the *consent fatigue* described in the literature and the need for novel automated consent methods in assistants [38]. However, and as one might expect, previous research [51] found that the more opportunities to learn the more accurate information sharing decisions, so it seems crucial to make the most of the very limited interactions one may have with a user to learn what their privacy preferences may be.

Recent user studies have actually focused on how users would like assistants to help them manage their privacy [11]. When it comes to the level of automation assistants should have, the study found similar evidence to previous studies [50], i.e., that users would like as much control as possible while intervening the least possible. In addition, this general finding had some specific nuances, where users would like to choose how much they will intervene and how much their privacy is managed automatically. The study also found that users should be given transparency about the decisions made and the opportunity to *review* the decisions made for auditing the decisions.

We take the evidence of these previous studies as requirements for the design of privacy-enhanced PAs [40, 42]. That is, PAs should manage users privacy in a way in which they should learn users' privacy preferences as much as possible from the user while minimizing the burden on the user, that users should be given a choice of how much they want to intervene, and that users should be given be a level of transparency for the decisions made, i.e., the model should be interpretable[1], as well as the opportunity to review the decisions made.

Based on these requirements, we present a novel model for privacy-enhanced PAs with two key mechanisms: i) a Dialogue Mechanism (DiM); and ii) a Decision making Mechanism (DeM). The DiM aims to understand user preferences and improve the performance of the PA with few interactions, by prioritizing the questions it poses to users. It also allows users to review PA's decisions so it can keep learning as it goes along. The DeM is a decision-making mechanism that is loosely based on a Case-based Reasoning (CBR)

* Corresponding Author. Email: xiao.zhan@kcl.ac.uk.

[1] Note interpretable means the opposite to black-box [36], that is, a model that is transparent about the decisions it takes and what they are based on. Note also the difference with *explainability*, which is also a desirable property, but out of the scope of this paper, as we do not focus on engineering the exact and best explanations that would be given to users (nor the specific social process needed for this [27]) based on an interpretable model, which is a related but different problem [35].

approach, where user and context similarity is used to pick the best decision for the current context and user (even if the context and the user are unknown). The DeM is interpretable as it can provide the most similar user and/or most similar context that led to a particular decision. We show experimentally using a dataset from a user study on privacy preferences for smart home PAs that the model performs substantially better than other online learning approaches with little user input, and particularly better than black-box alternatives.

2 Preliminaries

We follow the modern conceptualisation of privacy as dependent on the context according to Contextual Integrity (CI) [32], i.e. the same information flow may or may not lead to privacy violations, depending on the context. Contextual Integrity considers the following factors as determining the context that can make an information flow more or less appropriate: (1) the sender of the information, (2) the attribute or type of the information, (3) the subject of the information that is being transferred, (4) the recipient of the information, and (5) the transmission principles imposed on the transfer of the information from the sender to the recipient. Based on Contextual Integrity one can define and elicit privacy norms [8, 9], which are based on the *appropriateness* of information flows in a particular context.

Definition 1 (**Context**). *Given the set of Contextual Integrity parameters A, and, for each parameter $a \in A$, a domain D_a of values for the parameter, we define a context $c = \langle v_1, \ldots, v_k \rangle$, so that $v_i \in D_{a_i}$ and $k = |A|$.*

We exclude the parameter $sender$, as the sender is the PA, because our model is for a PA to decide on the information flows it originates. Since transmission principles in CI theory condition the flow of information from party to party, this may relate to several aspects of transmission.

Example 1. *Given the set of parameters $A = \{data, subject, recipient, purpose\}$, the context $c = \langle location, user, PA\_provider, security \rangle$ represents the case where the PA sends the user's location to the PA provider for security reasons (e.g. to ensure the PA is connecting from a known place).*

Definition 2 (**Privacy Norm**). *A privacy norm n is a tuple $\langle deontic, c \rangle$, where:*

- *$deontic$ represents the deontic modality, namely Obligation (O), Permission (P) and Prohibition (F).*
- *c is the context that the deontic modality applies to.*

Example 2. *Following the previous example, a privacy norm n that regulates that it is permitted to send the user's location to the PA provider for security reasons can be represented as:*

$$n = \langle P, \langle location, user, PA\_provider, security \rangle \rangle$$

3 Privacy Enhanced Model (PEM)

In this paper, we propose a privacy-enhanced model to help PAs reason about the best information-flow decision on different contexts, including known cases and those cases the current user has not experienced before. To achieve this aim, the model loosely follows a Case-based Reasoning [24] approach. The model has a knowledge base (KB) of norms for each user, which contains what contexts they would find appropriate for information flows to happen. This KB is used by the decision making mechanism (DeM) in order to, when a new context comes, *retrieve* and *reuse* (in CBR terminology) the best norm to deal with the context based on user and context similarity. The model also includes a dialogue mechanism (DiM), which allows the user to *revise* decisions made and, where pertinent, *retain* them in the KB (e.g. for when the PA is deployed the first time for a user). The DiM also allows for a very lightweight first dialogue with a new user not present in the KB. Next, we detail each of the components — KB, DeM and DiM, which are summarized in Figure 1.

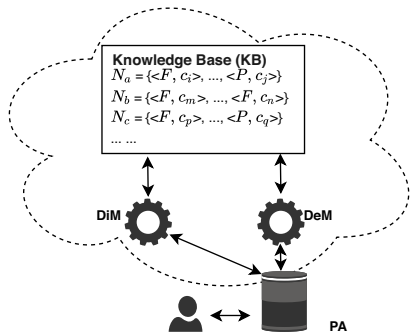

Figure 1. The components of the Model. The KB contains the previous cases. The DeM will make a decision once the PA needs to decide about a new context. The DiM is triggered to converse with the user and update the KB where pertinent.

3.1 Knowledge Base

Our model assumes a knowledge base that contains representations of users and privacy norms that can be used to reason out decisions. We formally define the knowledge base as follows.

Definition 3 (Knowledge Base). *Given a set of users U, the Knowledge Base is a set $KB = \{N_1, \ldots, N_m\}$, where N_u is the set of norms of user $u \in U$ and $m = |U|$.*

3.1.1 KB Initialization

We assume that the knowledge base starts with a set of other users and their norms before interacting with the current user. This can be achieved by a range of methods, such as product manufactures generating privacy norms that are based on user experience they have gathered in the reviews of PAs already put into use, during the market research phase, or based on user studies.

3.1.2 KB Update

When there are new norms to be added to the KB (based on the DiM as explained later), the model checks whether the new norms to be added conflict with a norm that already exists in the KB. In this paper, we assume that a conflict arises when an action is simultaneously prohibited and permitted/obliged, and its variables have overlapping values.

Definition 4 (Norm Conflict). *Given two norms $n = \langle Deontic_n, c_n \rangle$ and $m = \langle Deontic_m, c_m \rangle$. We say they are in conflict, denoted as $conflict(n, m)$ iff*

$$Deontic_n \neq Deontic_m \wedge c_n = c_m.$$

For instance, a conflict occurs between a prohibition and a permission if the for the same context. In this paper, as we follow an online learning approach, that is, we want to be aligned with the user as we learned from them, the most recent norm takes precedence over the older norm if they are in conflict.

3.2 Decision Making Mechanism

The Decision making mechanism (DeM) is at the heart of the model's ability to guide the PA's responses when it detects a context that requires a decision. Once a context is to be considered by the model, the DeM will be triggered to reason out a decision on whether the action associated should be allowed to occur. Our model adapted the four-step procedure (also known as the 4R cycle) of CBR [1], to execute the whole decision-making process. The four-step procedure consists of the steps *retrieve, reuse, revise, and retain*, which is identified as a proper way to apply CBR to an application [1].

3.2.1 Case Similarity

The decision-making process is inextricably linked to two similarity functions. One is used to compute the similarity between two **contexts** (sc), and the other is to compute the similarity between two **users** (su).

First, we start with the context similarity, which computes how similar two contexts are by looking at how similar their parameters are in turn.

Definition 5 (Context Similarity)**.** *Given two contexts c and d, their similarity is:*

$$sc(c,d) = \sum_{j=1}^{k} w_j \times sim_j(c.v_i, d.v_i) \tag{1}$$

where k is, as before, the number of CI parameters, and w_j is the weight of the j-th parameter, which represents the importance of parameter j, where $\sum_{i=1}^{k} w_i = 1$. The function $sim_j(\cdot,\cdot)$ represents the similarity between each pair of parameters. Note that $sim_j(\cdot,\cdot)$ may be different depending on the parameter.

For instance, in this paper, in our experimental setup, we defined $sim_j(\cdot,\cdot)$ based on empirical evidence of users. In particular, regarding similarity $sim_j(\cdot,\cdot)$, we defined it as follows: 1) data type similarity depends on the sensitivity level of data, so that the more similar their sensitivity the more similar the data types are considered; 2) recipient similarity depends on the relationship between the recipient and the user, so the more similar the relationship the more similar the recipient; 3) the similarity of the rest of the parameters simply looks at how many of them are the same. Further details can be found in the experimental section.

Regarding the weights of each parameter similarity w_k, one could automatically calculate that from the KB. For instance, one way to do this is as we did for our experiments, where, using the KB, we construct a regression model that considers as independent variables the parameters that define contexts and then the deontic modality as the independent variable. The coefficients of the regression model for each parameter could then be used to set the different weights w_k. Further details can be found in the experimental section.

Next, we focus on the similarity between users, which depends on the similarity between their norms (which, in turn, also depend on context similarity as below).

Definition 6 (User Similarity)**.** *Given two users i and j, their similarity is:*

$$su(i,j) = \frac{O_{i,j}}{|N_i|} + \left(1 - \frac{O_{i,j}}{|N_i|}\right) * \frac{L_{i,j}}{|N_i| - O_{i,j}} \tag{2}$$

where $O_{i,j}$ is the number of privacy norms of users i and j that are the same, i.e., that have the same deontic modality and the same scenario; and $L_{i,j}$ is the number of norms of users i and j that are not the same but that have the most similar context among all of the norms of i and j and have the same deontic modality. Note that this is done from the point of view of user i.

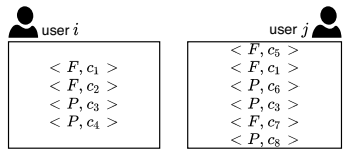

Figure 2. An example of norms for two users to show how the model calculates the similarity between user i and j using function $su(\cdot)$ (see Example 3).

Example 3. *For users and their privacy norms shown in Fig. 2, we have that $O_{i,j} = 2$, as there are two norms of them are the same — i.e., they have the same modality for the same context. For the remaining norms in N_i, c_5 is the most similar to c_2. c_8 is the most similar to c_4. For these two pairs, only one has the same deontic modality, which is F, therefore $L = 1$. So the final similarity $sc(i,j) = 0.75$.*

3.2.2 Decision Making Process

We now focus on how the model makes a decision based on user and context similarity. The steps to make a decision about a context c is as follows — and as described in pseudocode in Algorithm 1:

Step 1) Same context from user or most similar users. In the first step, the model tries to find whether any of the norms of the user or the μ most similar users has that same context c (Algorithm 1 Lines 1–5). Note that the number of most similar users μ is a paramater of the model, which means how many of the most similar users to the user will be considered. The similarity between the user and other users is calculated using Definition 6 and updated when the DiM runs as explained later on in Section 3.3. Note also that users are considered in order of similarity, that is, the current user goes first, then the most similar user to the current user, and so on, so the algorithm stops as soon as the same context as c is found in user similarity order. If the c context is found in one of the norms of the user or their μ most similar users, then the deontic modality of the norm that contains it is returned. If none of the norms of the user or the μ most similar users relates to context c, then the next step is to find the most similar context.

Step 2) Most similar context. As in the previous step, the norms of the user and the μ most similar users are considered again, but this time the model looks for the norm that has the context with the highest similarity to c (Algorithm 1 Lines 7–15). When it finds the norm with the most similar context to c, then the model considers how high that similarity is with respect to the context similarity threshold θ. If the similarity between the context of the norm found and c is higher or equal that the threshold θ, then the deontic modality returned will be that of the norm found. This is to ensure that there is a minimum context similarity — note that in the experimental section we consider different values of θ and how they influence the performance

of the model. Otherwise, if the similarity between the context of the norm found and c is lower than the threshold θ, then the model considers all the norms with a context that is most similar to c for the user and their μ most similar users. That is, the model ends up with one norm per user — the one of that user that is most similar to c. Then, it takes a majority vote considering the deontic modality of each of these norms, so that the deontic modality that is the majority is then returned as the decision to be taken (Algorithm 1 Lines 16–23).

Algorithm 1 Decision-making Process

Require: context c, knowledge base KB, main user a, array of most μ similar users and a ordered by similarity $users$, context similarity thresthold θ
Ensure: Decision on c
1: **for each** $i = 0$ to μ **do**
2: **if** $\exists n \in N_{users[i]}$ s.t. $n.c = c$ **then**
3: **return** $n.deontic$
4: **end if**
5: **end for**
6: $max \leftarrow 0$
7: **for each** $i = 0$ to μ **do**
8: **for each** $n \in N_{users[i]}$ **do**
9: $sim = sc(c, n.c)$
10: **if** $sim > max$ **then**
11: $max \leftarrow sim$
12: $deontic \leftarrow n.deontic$
13: **end if**
14: **end for**
15: **end for**
16: **if** $max < \theta$ **then**
17: **for each** $i = 0$ to μ **do**
18: $aux \leftarrow \underset{n \in N_{users[i]}}{\arg\max}\ sc(c, n.c)$
19: $d[i] \leftarrow aux.deontic$
20: **end for**
21: $deontic \leftarrow MajorityVoting(d)$
22: **end if**
23: **return** $deontic$

3.3 Dialogue Mechanism

When the model is to be used with a new user that is not in the KB - e.g. when the PA is deployed after being purchased, it is infeasible for the model to ask about every possible context, or even many contexts, as this would be overwhelming for the user. At the same time, the model needs a minimum in order to be able to compute the similarity of the current user with users in the KB. Therefore the model incorporates a dialogue mechanism (DiM) that can interact with the user for two main purposes: 1) help the PA to initialise to the current user, and 2) let the user review the decisions made by the model. In this section, we will detail how these processes work and how the model will update its KB accordingly.

3.3.1 New User Initial Norms

The model is initialized when first deployed. The purpose of this is to gain an initial understanding of the user and to set the stage for the subsequent reasoning process but minimising the information required from the user. Here, it is crucial to get as much information with as less intervention demanded from the user. In particular, asking about many contexts or context parameters would make it a burden on the user. Instead, we focus on the parameters and their values that can play a bigger role in ascertaining the similarity of the new user with already existing users in the KB.

In order to do this, the model establishes an order between the parameters and the values of the parameters. Specifically:

Definition 7 (Parameter order)**.** *Given the set of contextual integrity parameters A, a parameter order is a partial order $\preceq_A$, so that $(A, \preceq_A)$ is a partially ordered set.*

In practice, $\prec_A$ can take different forms. One possible approach, as we use in our experimental section, is to define the partial order $\preceq_A$ based on the *influence* exerted by each parameter in the set A on the acceptability of the associated contexts. To achieve this, we can employ statistical methods such as logistic regression, which allows us to examine the relationship between the coefficients corresponding to each parameter in an interpretable manner.

Definition 8 (Value order)**.** *Given a contextual integrity parameter $a \in A$ with domain $D = \{d_1, \ldots, d_n\}$, a value order is a partial order $\preceq_D$, so that $(D, \preceq_D)$ is a partially ordered set.*

In practice, $\preceq_D$ can also take different forms. In this case, we focus on the values that seem to generate the most different decisions in the KB. That is, the ones which may inform the model the most to find similar users to the new user. We consider three different approaches to this, and compare them experimentally later: i) the values of the parameter that leads to the most diverse set of decisions in the KB - i.e., the ones where the differences between the users may be more apparent and which can help the most to find similar/dissimilar users; ii) the values of the parameter that are intrinsically known to be different in practice (e.g. for data types, those that are perceived as most or least sensitive), so that knowning about them from the new user can also help in finding the most similar users in the KB; iii) random ordering, which we mainly use as a baseline.

Once the parameter and value orders are established, then the model can enquire the user for a limited number of parameters, ω_A, and a limited number of values, ω_D, in the order established by $\preceq_A$ and $\preceq_D$. The responses are then used to create a set of norms that would be added to the KB using the KB update process described in 3.1.

Example 4. *Assuming that $\omega_A = 2$, $\omega_D = 2$, data_type $\preceq_A$ recipient, and music $\preceq_D$ banking_details, then the model would ask the user:*
Q1: "With whom would you share banking details?"
Q2: "With whom would you share music?"

Finally, the last step of the initilization is to compute the μ most similar users to the new user. This is done after the KB is updated with the new norms created through the initialization and based on the user similarity (Definition 6).

3.3.2 Decision Review

In addition to eliciting a number of initial norms with limited interactions with the user, the DiM would also let the user review the decisions the model has made. This is done at the frequency ϕ, which can be selected by the user — note that we show in the experiments the effect this frequency has on the quality of predictions made. The reviewed decisions will then be added to the KB. If the user is not satisfied with a particular case, the model will create a corresponding new norm and add it to the KB too, which will be updated as shown in the previous section. Finally, it is important to mention that after the review has happened, then the model recalculates the μ most similar users based on the new norms added.

4 Evaluation

In this section, we describe the procedure we used to evaluate the performance of our proposed model for privacy-enhanced personal assistants PEM, the influence of the different parameters it has on its performance, and how PEM compares to previous approaches form the literature.

4.1 Dataset

We use a fully-anonymized and publicly-available dataset[2] of real privacy decisions, which was the result of a survey of PA users in households [3]. The survey used a combination of various contextual integrity parameters, including 15 data types, 15 types of recipients (contains both internal and external recipients in the household), and 7 transmission principles, to create over a thousand different information flow contexts.

The total number of participants was 1,739. Each of the participants in the study was randomly assigned around 180 different information flow contexts. For each of the information flow contexts, each participants was asked about the acceptability of that flow. Therefore, the dataset contains a total of **292,478 decisions**.

The dataset contains five types of relationships for the 15 different types of recipients that we use for the recipient similarity (closest: partner, parents, children; close: siblings, close family, housemates; general user: visitor, housekeeper, visiting friends, neighbours; relevant parties: PA provider, third-party skills; other parties: advertising agencies, law enforcement agencies).

The dataset also contains the *sensitivity* (with 1 being the least sensitive and 5 being the most sensitive) of the data type as perceived by the user. We used this sensitivity straightaway to compute the similarity between data types, and for one of the approaches for the initialization step of the DiM (the one where the partial order of data types is defined according to their sensitivity).

Finally, the w_k weights used for context similarity for each different parameter are set based on the regression analysis made in the original publication of the dataset [3]. In particular, a ratio of importance of 5:3:2 for recipient, data type and other parameters, respectively, is derived from the coefficients of the regression model reported on the paper.

4.2 Parameter Influence

Two significant parameters need to be determined for our model, namely the number of similar users the model should retrieve (μ), and the frequency at which the user chooses to review the decisions made (ϕ). Examining the impact of fine-tuning the μ parameter on the model's performance can give us insight into whether a comprehensive KB, especially known extensive privacy norms of other users, is essential for making appropriate privacy decisions as a PA. Additionally, the effect of parameter ϕ on the model will indicate the relationship of user participation with the model's accuracy.

In addition, regarding the context similarity calculated in the decision making mechanism, we set a threshold θ to check whether it reaches sufficient similarity to make the final decision or instead using the majority vote algorithm. We decided to use 0.6, because: 1) according to the calculation of context similarity, a number above 0.6 means that at least the $data$ and $recipients$ described in the context are highly similar, 2) we also varied θ for some of the best combinations of ϕ and μ as shown later in this section. Regarding the initialization, we follow the approach stated in Section 3.3.1, where the partial order between parameters, $\preceq_A$, is set based on the coefficients of logistic regression, that is, as stated in the previous section for the w_k weights. For $\preceq_D$, which is the partial order between parameter values, we take the best approach of all the three considered, which is just based on the sensitivity of the data — we show later in this section how this fairs with respect to the other two alternatives. Finally, and to minimize the need for user intervention, we set both ω_A, ω_D to two, i.e., the model would only ask the user two questions (see example in Section 3.3.1).

We start by randomly splitting the users in a ratio of 1:9, where the 10% (180 users in the whole dataset) are regarded as new users that adopt PAs. They were evaluated on a one-by-one basis. Each experiment was repeated 10 times to validate our model on 10 different and completely random splits. The remaining 90% (n = 1559) users' data are regarded as the previous users, we create privacy norms of each user and store them in the KB as the default.

Table 1 shows the average accuracy of each model in different parameter combinations. As can be seen from the results in Table 1, by adjusting the number of similar users required and the frequency with which users are asked to review the decisions made by the model, the accuracy varies accordingly. When the number of similar users that the model should find (μ) is fixed, it is clear that the accuracy of the model is higher when users' views start to be involved ($\phi = 0.2$) than when the model relies only on context similarity calculations to obtain the results ($\phi = 0$). In addition, the accuracy of the model improved as user involvement increased. In addition, the model achieved the highest accuracy if users reviewed decisions most frequently (every time). However, user reviews did not improve too much on the accuracy of the model, for example, the accuracy when $\phi = 1$ is at most 0.042 better than when $\phi = 0$. This means that the DiM initialization step is highly effective in eliciting a small number of norms that then help in terms of finding similar users and contexts in the KB that are useful for an accurate decision.

An interesting finding is that, for the same review frequency (ϕ), the results don't seem to always improve as the number of similar users increases (from $\mu = 1$ to $\mu = 20$). The model achieves the best results for all review frequencies when μ is between 5 and 10. This suggests that it is not necessary for the model to spend a lot of time in finding a large number of similar users to use as references for making decisions, i.e. only a few very similar users are enough to get a good result.

Coming now back to the other parameters, Table 2 shows the influence of the parameter θ, that is the threshold used for whether contexts are similar enough. It can be seen in the table that the influence of θ for some combinations of ϕ and μ confirms that a value of $\theta = 0.6$ is a good choice.

Regarding the three potential approaches for initializing the model in Section 3.3.1, Table 3 shows the results when: i) only asking about the parameters that lead to the most diverse set of decisions in the KB - this is done by computing the standard deviation, so that higher standard deviation is indicative of a greater dispersion in the data (for this a different value for the number of contexts from the KB is selected and represented by n in the Table); ii) asking about values that are perceived differently by users, that are the most sensitive and the least sensitive data types; and iii), randomly picking two data types. The results confirm that prioritizing based on data sensitivity works the best in this case.

[2] The dataset is publicly available from here: https://osf.io/63wsm/.

Table 1. Model accuracy with different parameter combinations. As mentioned in the Section 4.2, each experiment was repeated 10 times to validate the model performance. Result of accuracy are rounded to three decimal places, and numbers in *brackets* are the standard deviation of the 10 folds. As the number of information flows rated by each user in the original dataset is different, the number of contexts used for testing during the experiment varies, e.g. user 1 has 156 data (context) to make a decision on, while user 3 has 170 data. In the table, we have therefore also used an "≈" to indicate the approximate number of data to be reviewed, rather than a definite number, i.e, around every 64 decisions when $\phi = 0.4$.

		Frequency					
		$\phi = 0$ [No review]	$\phi = 0.2$ [≈Every 128]	$\phi = 0.4$ [≈Every 64]	$\phi = 0.6$ [≈ Every 32]	$\phi = 0.8$ [≈Every 16]	$\phi = 1$ [Every time]
No. of sim users	μ =1	0.783 (0.05)	0.787 (0.08)	0.792 (0.06)	0.796 (0.04)	0.798 (0.08)	0.825 (0.06)
	μ =5	0.835 (0.04)	0.838 (0.06)	0.840 (0.06)	0.842 (0.05)	0.849 (0.03)	0.851 (0.06)
	μ =10	0.814 (0.04)	0.816 (0.06)	0.825 (0.07)	0.826 (0.06)	0.830 (0.05)	0.834 (0.04)
	μ =15	0.818 (0.05)	0.819 (0.05)	0.821 (0.06)	0.827 (0.04)	0.833 (0.08)	0.835 (0.05)
	μ = 20	0.811 (0.04)	0.814 (0.04)	0.818 (0.04)	0.824 (0.04)	0.827 (0.06)	0.830 (0.06)

Table 2. Experiments varying θ for combinations of μ and ϕ.

		θ										
		0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Models	μ=5 ϕ=0	0.790	0.796	0.796	0.798	0.802	0.825	0.835	0.835	0.831	0.826	0.820
	μ=5 ϕ=0.4	0.800	0.821	0.830	0.832	0.835	0.838	0.840	0.837	0.835	0.835	0.832
	μ=5 ϕ=0.8	0.826	0.828	0.837	0.840	0.844	0.849	0.849	0.848	0.846	0.846	0.845

Table 3. Initialization approaches comparison (n is # cases from KB).

	i)					ii)	iii)
	n =5	n=10	n=15	n=20	n=30		
μ=5 ϕ=0	0.736	0.752	0.798	0.802	0.820	0.835	0.732
μ=5 ϕ=0.4	0.752	0.784	0.812	0.816	0.832	0.840	0.740
μ=5 ϕ=0.8	0.766	0.810	0.822	0.825	0.837	0.849	0.753

4.3 Comparison with Other Approaches

Next, we compare our model with other interpretable and non-interpretable approaches. We also consider a baseline, control condition where the decision is random.

Regarding interpretable approaches, we compare with [51], which uses a hybrid learning mechanism combining data mining to mine a set of general rules that match the behavioural norms of most users, and then reasons out the final decision based on these general norms and user-specific feedback.

We also consider online learning methods that have a degree of interpretability, such as the online learning versions of decision tree and logistic regression. Online learning is a method of machine learning for data arriving in a *sequential* order, where a learner aims to learn and update the best predictor for future data at every step. We chose online learning because it can also start with an initial set of data and then update the model as it gets more instances (e.g. by a new user or after review), so it can therefore compare with what our model does. Specifically, we considered two online learning approaches: Very Fast Decision Tree VFDT [18] with majority class in leaves for classification, information gain as the heuristic measure, and default parameters δ=10^{-7}, $nmin$=200, $\tau = 0.05$; and an incremental logistic regression model using River [28] with the default learning rate of 0.05 and a window size of 50.

Finally, we also compare with a black-box neural network approach. In particular, The Multilayer Perceptron (MLP) was built where each attribute (datatype, recipient, purpose, and condition) is first passed through its respective embedding layer to get 64-dimensional embeddings. These embeddings were passed to 2 hidden layers with 128 hidden units. The output layer is a Softmax layer for classification with 2 classes (*'Acceptable'* and *'Unacceptable'*). The optimizer used is the Stochastic Gradient Descent (SGD) optimizer.

For our model, we use the following combinations of parameters. The first $PEM1$, with $\mu = 5$, $\phi = 0.8$; the second $PEM2$, with $\mu = 5$, $\phi = 0.4$; and the third $PEM3$, with $\mu = 5$, $\phi = 0$. The rest of parameters we leave them as in the previous section. This is because the result of these parameters combinations give a relative higher accuracy according to the experiments in the previous section, and they are also representative of different review needs, because, given that the process of reviewing decisions is carried out through the dialogue mechanism, reviewing too many cases per time would put extra burden on the user that they would face regularly from time to time (as they can specify with ϕ).

Table 4. Performance comparison with other approaches.

Model	Interpretable	Accuracy
$PEM1$: Review every 16 cases	✓	0.849
$PEM2$: Review every 64 cases	✓	0.840
$PEM3$: No Review	✓	0.835
RIVER incremental learning [28]	✓	0.772
Zhan et al. [51]	✓	0.741
Very fast decision tree (VFDT) [18]	✓	0.706
Neural network (MLP)	✗	0.680
Baseline (Random decision)	✗	0.501

The result comparing the performance of different versions of PEM with previous approaches in the literature can be seen in Table 4. As expected, the baseline, random decision approach shows an accuracy close to 0.5, and it is the worst of all the approaches tried. When it comes to the other approaches, PEM, regardless of the version is shows the best performance, with the added benefit of being interpretable. Interestingly, PEM works better than the other approaches we compared it with even in the case where the user would not review any of the decisions made. This suggests that the initialization step of the DiM is highly effective, that is, with only two questions asked to the user, it can effectively find other similar users that can help then the DeM make very accurate predictions. One can

also see that online machine learning approaches seem to work better for this case than neural networks, which could be due to the dynamic nature of the problem as well as to the fact that neural networks usually require a huge amount of data for accurate results, which, as in this case, may not always be available.

5 Related Work

The burgeoning demand for privacy-preserving models that safeguard user data from unauthorized access has become a paramount concern in contemporary data-driven societies [13, 26]. A particularly promising approach to addressing this exigency entails the development of sophisticated models capable of discerning individual privacy predilections and subsequently inferring privacy decisions in line with user expectations.

Previous researchers have primarily focused on leveraging extensive datasets and machine learning algorithms to devise classifiers adept at predicting privacy decisions [39, 45]. These classifiers are frequently trained on textual or visual features, employing copious amounts of user-labeled data to accurately predict categorizations for test set data. Beyond machine learning, other more symbolic approaches, such as agent-based models, have been proposed for helping manage privacy in social networks, from those focusing on individual privacy recommendations and policy violations [25, 20, 22] to those cases where many users are involved [21, 48, 46, 17, 30].

As the application scenario shifts from sharing photos on social networks to the interaction between users and personal assistants, and even as users manage their privacy needs across multiple devices, *context* has become a focus for researchers [33, 19, 4]. This notion serves as the impetus for Kokciyan et al. [23] to put forth a situation-based model for privacy protection. The model examines diverse contexts using natural language processing algorithms and SVM classifiers to categorize contexts and scenarios. This approach also utilizes user trust ratings for the scenarios and it appears to yield a good result in terms of accuracy. Yet, it necessitates the trust ratings provided by users for a considerable number of scenarios initially, which may not always be available, for instance in the dataset used in this paper. It is also worth noting that the model's predictions are contingent upon the trustworthiness of the context, rather than identifying the context that the user prefers to trust more or less. Furthermore, owing to the employment of techniques like sentence embeddings and SVM, the reasons behind their model's decisions may not always be easily traceable nor interpretable.

In another study related to our research, Amoros et al. [7] put forth a method that utilizes a collaborative filtering technique to predict user preferences in terms of the degree of acceptability of a particular information flow. It is crucial to note, however, that their focus lies on predicting this specific degree of acceptability rather than a 'binary decision' of whether the assistant should share the information or not. Therefore, the model may be useful for predicting preferences but not directly for a PA to be able to make a decision about whether data should be shared or not. This also means that it was impossible for us to compare with this work, as we could not use their model and implementation for our case study. In addition, the method proposed in [7] does not consider aspects to optimize the initialization beyond considering random questions to pose to the user, the input from the user is considerable (they need information about many contexts), which may become a burden on the user, and there is not a notion of transparency or review of any decisions as they only focus on predicting preferences.

6 Conclusion

In this paper we presented a privacy-enhanced model for personal assistants. The model has two main components, a dialogue mechanism and a decision making mechanism, that allow it to learn the best information sharing decisions aligned with users' privacy preferences with minimal user intervention. It also offers a degree of transparency by being interpretable and allowing the revision of the decisions made. We showed experimentally that the model performs considerably better than previous works.

In the future, we would like to build on the model's interpretability to engineer and automatically generate *explanations* for the reasons of a specific decision made, as privacy explanations for assistants in particular have been shown to assuage privacy concerns [37]. While the model is interpretable and one can inspect it to come up with the most similar user or context that inspired the current decision as explained above, the problem of deciding exactly what information (e.g. about the user or the context) is presented to users, how this information is presented, and the procedure to engage in a query/answer dialogue with the user to explain the decisions as a social process, as suggested in [27]) for designing explainable AI , is an exciting but non-trivial problem. One such possible approach that could be used as an interesting starting point for explainability is to apply argumentation, as done in [35, 49, 29, 31] for explaining recommendation systems and case-based reasoning systems. Finally, we focused on single user preferences, but future work could use them as input to existing multiuser privacy models [30, 43, 41] or norm-based access control with multiple users [47, 12].

Acknowledgements

This research was supported by UKRI through REPHRAIN (EP/V011189/1), the UK's Research centre on Privacy, Harm Reduction and Adversarial Influence online, as part of its PRAISE project. Xiao Zhan is funded by King's PGR International Scholarship.

References

[1] Agnar Aamodt and Enric Plaza, 'Case-based reasoning: Foundational issues, methodological variations, and system approaches', *AI communications*, **7**(1), 39–59, (1994).
[2] Noura Abdi, Kopo Ramokapane, and Jose Such, 'More than smart speakers: security and privacy perceptions of smart home personal assistants', in *Fifteenth USENIX Symposium on Usable Privacy and Security ({SOUPS} 2019)*, (2019).
[3] Noura Abdi, Xiao Zhan, Kopo M Ramokapane, and Jose Such, 'Privacy norms for smart home personal assistants', in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–14, (2021).
[4] Nirav Ajmeri, Hui Guo, Pradeep K Murukannaiah, and Munindar P Singh, 'Robust norm emergence by revealing and reasoning about context: Socially intelligent agents for enhancing privacy', in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pp. 28–34. IJCAI, (2018).
[5] Amazon. Alexa skills. Video, March 2008.
[6] Tawfiq Ammari, Jofish Kaye, Janice Y Tsai, and Frank Bentley, 'Music, search, and iot: How people (really) use voice assistants.', *ACM Trans. Comput. Hum. Interact.*, **26**(3), 17–1, (2019).
[7] Marc Serramia Amoros, William Seymour, Natalia Criado, and Michael Luck, 'Predicting privacy preferences for smart devices as norms', in *The 22nd International Conference on Autonomous Agents and Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS), (2023).
[8] Noah Apthorpe, Yan Shvartzshnaider, Arunesh Mathur, Dillon Reisman, and Nick Feamster, 'Discovering smart home internet of things privacy norms using contextual integrity', *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, **2**(2), 59, (2018).

[9] Noah Apthorpe, Sarah Varghese, and Nick Feamster, 'Evaluating the contextual integrity of privacy regulation: Parents' iot toy privacy norms versus {COPPA}', in *28th {USENIX} Security Symposium ({USENIX} Security 19)*, pp. 123–140, (2019).

[10] Mary K. Bispham, Clara Zard, Suliman Sattar, Xavier Ferrer Aran, Guillermo Suarez-Tangil, and Jose Such, 'Leakage of sensitive information to third-party voice applications', in *ACM CUI 2022: 4th Conference on Conversational User Interfaces*, pp. 32:1–32:4, (2022).

[11] Jessica Colnago, Yuanyuan Feng, Tharangini Palanivel, Sarah Pearman, Megan Ung, Alessandro Acquisti, Lorrie Faith Cranor, and Norman Sadeh, 'Informing the design of a personalized privacy assistant for the internet of things', in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–13, (2020).

[12] Natalia Criado and Jose Such, 'Implicit contextual integrity in online social networks', *Information Sciences*, **325**, 48–69, (2015).

[13] Daniel J Dubois, Roman Kolcun, Anna Maria Mandalari, Muhammad Talha Paracha, David Choffnes, and Hamed Haddadi, 'When speakers are all ears: Characterizing misactivations of iot smart speakers', *PoPETS*, **2020**(4), 255–276, (2020).

[14] Jide Edu, Xavier Ferrer-Aran, Jose Such, and Guillermo Suarez-Tangil, 'Measuring alexa skill privacy practices across three years', in *Proceedings of the Web Conference (WWW)*, (2022).

[15] Jide Edu, Xavier Ferrer-Aran, Jose Such, and Guillermo Suarez-Tangil, 'Skillvet: Automated traceability analysis of amazon alexa skills', *IEEE Transactions on Dependable and Secure Computing (TDSC)*, **20**(1), 161–175, (2023).

[16] Jide Edu, Jose Such, and Guillermo Suarez-Tangil, 'Smart home personal assistants: a security and privacy review', *ACM Computing Surveys (CSUR)*, **53**(6), 1–36, (2020).

[17] Ricard L Fogues, Pradeep K Murukannaiah, Jose Such, and Munindar P Singh, 'Sharing policies in multiuser privacy scenarios: Incorporating context, preferences, and arguments in decision making', *ACM Transactions on Computer-Human Interaction (TOCHI)*, **24**(1), 1–29, (2017).

[18] Geoff Hulten, Laurie Spencer, and Pedro Domingos, 'Mining time-changing data streams', in *Procs of the ACM SIGKDD conference on Knowledge discovery and data mining*, pp. 97–106, (2001).

[19] Pramod Jagtap, Anupam Joshi, Tim Finin, and Laura Zavala, 'Preserving privacy in context-aware systems', in *2011 IEEE Fifth International Conference on Semantic Computing*, pp. 149–153. IEEE, (2011).

[20] Özgür Kafali, Akin Günay, and Pınar Yolum, 'Protoss: A run time tool for detecting privacy violations in online social networks', in *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 429–433. IEEE, (2012).

[21] Nadin Kökciyan, Nefise Yaglikci, and Pınar Yolum, 'An argumentation approach for resolving privacy disputes in online social networks', *ACM Transactions on Internet Technology (TOIT)*, **17**(3), 1–22, (2017).

[22] Nadin Kökciyan and Pınar Yolum, 'Priguard: A semantic approach to detect privacy violations in online social networks', *IEEE Transactions on Knowledge and Data Engineering*, **28**(10), 2724–2737, (2016).

[23] Nadin Kökciyan and Pınar Yolum, 'Taking situation-based privacy decisions: Privacy assistants working with humans', in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pp. 703–709, (2022).

[24] Janet Kolodner, *Case-based reasoning*, Morgan Kaufmann, 2014.

[25] A Can Kurtan and Pınar Yolum, 'Assisting humans in privacy management: an agent-based approach', *Autonomous Agents and Multi-Agent Systems*, **35**(1), 1–33, (2021).

[26] Abraham Mhaidli, Manikandan Kandadai Venkatesh, Yixin Zou, and Florian Schaub, 'Listen only when spoken to: Interpersonal communication cues as smart speaker privacy controls', *Proceedings on Privacy Enhancing Technologies*, **2020**(2), 251–270, (2020).

[27] Tim Miller, 'Explanation in artificial intelligence: Insights from the social sciences', *Artificial intelligence*, **267**, 1–38, (2019).

[28] Jacob Montiel, Max Halford, Saulo Martiello Mastelini, Geoffrey Bolmier, Raphael Sourty, Robin Vaysse, Adil Zouitine, Heitor Murilo Gomes, Jesse Read, Talel Abdessalem, et al., 'River: machine learning for streaming data in python', (2021).

[29] Francesca Mosca, Ştefan Sarkadi, Jose M Such, and Peter McBurney, 'Agent expri: Licence to explain', in *International workshop on explainable, transparent autonomous agents and multi-agent systems*, pp. 21–38. Springer, (2020).

[30] Francesca Mosca and Jose Such, 'Elvira: an explainable agent for value and utility-driven multiuser privacy', in *International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pp. 916–924, (2021).

[31] Francesca Mosca and Jose Such, 'An explainable assistant for multiuser privacy', *Autonomous Agents and Multi-Agent Systems (JAAMAS)*, **36**(10), 1–45, (2022).

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

[33] Gideon Ogunniye and Nadin Kökciyan, 'Argumentation-based dialogues for privacy policy reasoning', in *The 3rd Annual Symposium on Applications of Contextual Integrity*, (2021).

[34] Jay Patel, Sundar Anand, and Rohan Luthra, 'Image-based smart surveillance and remote door lock switching system for homes', *Procedia Computer Science*, **165**, 624–630, (2019).

[35] Antonio Rago, Oana Cocarascu, Christos Bechlivanidis, David Lagnado, and Francesca Toni, 'Argumentative explanations for interactive recommendations', *Artificial Intelligence*, **296**, 103506, (2021).

[36] Cynthia Rudin, 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead', *Nature machine intelligence*, **1**(5), 206–215, (2019).

[37] William Seymour, Mark Cote, and Jose Such, 'Ignorance is bliss? the effect of explanations on perceptions of voice assistants', in *PACM on Human-Computer Interaction - ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*, (2023).

[38] William Seymour, Mark Coté, and Jose Such, 'Legal obligation and ethical best practice: Towards meaningful verbal consent for voice assistants', in *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, pp. 166:1–166:16, (2023).

[39] Anna Squicciarini, Cornelia Caragea, and Rahul Balakavi, 'Toward automated online photo privacy', *ACM Transactions on the Web (TWEB)*, **11**(1), 1–29, (2017).

[40] Jose Such, 'Privacy and autonomous systems', in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pp. 4761–4767, (2017).

[41] Jose Such and Natalia Criado, 'Resolving multi-party privacy conflicts in social media', *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, **28**(7), 1851–1863, (2016).

[42] Jose Such, Agustin Espinosa, and Ana Garcia-Fornes, 'A survey of privacy in multi-agent systems', *The Knowledge Engineering Review*, **29**(03), 314–344, (2014).

[43] Jose Such and Michael Rovatsos, 'Privacy policy negotiation in social media', *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, **11**(1), 4, (2016).

[44] Madiha Tabassum, Tomasz Kosinski, and Heather Richter Lipford, '" i don't own the data": End user perceptions of smart home device data practices and risks', in *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*, pp. 435–450, (2019).

[45] Lam Tran, Deguang Kong, Hongxia Jin, and Ji Liu, 'Privacy-cnh: A framework to detect photo privacy with convolutional neural network using hierarchical features', in *Thirtieth AAAI Conference on Artificial Intelligence*, (2016).

[46] Onuralp Ulusoy and Pınar Yolum, 'Norm-based access control', in *Proceedings of the 25th ACM Symposium on Access Control Models and Technologies*, SACMAT '20, p. 35–46, New York, NY, USA, (2020). Association for Computing Machinery.

[47] Onuralp Ulusoy and Pınar Yolum, 'Norm-based access control', in *Proceedings of the 25th ACM symposium on access control models and technologies*, pp. 35–46, (2020).

[48] Onuralp Ulusoy and Pınar Yolum, 'Panola: A personal assistant for supporting users in preserving privacy', *ACM Trans. Internet Technol.*, **22**(1), (sep 2021).

[49] Wijnand van Woerkom, Davide Grossi, Henry Prakken, Bart Verheij, Kristijonas Čyras, Timotheus Kampik, Oana Cocarascu, Antonio Rago, et al., 'Justification in case-based reasoning', in *Proceedings of the First International Workshop on Argumentation for eXplainable AI*, pp. 1–13. CEUR Workshop Proceedings, (2022).

[50] Eric Zeng and Franziska Roesner, 'Understanding and improving security and privacy in multi-user smart homes: a design exploration and in-home user study', in *28th {USENIX} Security Symposium ({USENIX} Security 19)*, pp. 159–176, (2019).

[51] Xiao Zhan, Stefan Sarkadi, Natalia Criado, and Jose Such, 'A model for governing information sharing in smart assistants', in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 845–855, (2022).