# Extended Abstract of Privacy-enhanced AI Assistants based on Dialogues and Case Similarity

Xiao Zhan
King's College London
UK
xiao.zhan@kcl.ac.uk

Stefan Sarkadi
King's College London
UK
stefan.sarkadi@kcl.ac.uk

Jose Such
Universitat Politecnica de Valencia
Spain
jsuch@upv.es

***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. 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.**

***Index Terms*—Jornadas, Ciberseguridad**

**Tipo de contribución:** *Investigación ya publicada (límite 2 páginas)*

## I. 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 [1], [2], [3]. PAs have a distinct working ecosystem of their own, which is complicated and involves many different stakeholders [1], [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 [4], [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 [6].

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 [1]. 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 [7], those mechanisms end up not being utilized in practice because users find it too burdensome [7]. 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 [7]. 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 [8]. However, and as one might expect, previous research [9] 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 [10]. When it comes to the level of automation assistants should have, the study found similar evidence to previous studies [7], 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 PAsThat 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, as well as the opportunity to review the decisions made.

Based on these requirements, we present in [11] 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) 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.

## II. Privacy Enhanced Model (PEM)

In [11], 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 [12] 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. As we show experimentally in [11], with only two questions asked to the user through the DiM, this allows the PEM to produce very accurate predictions. The detail each of the components (KB, DeM and DiM), which are summarized diagramatically in Figure 1 below, can be found in the original paper [11].

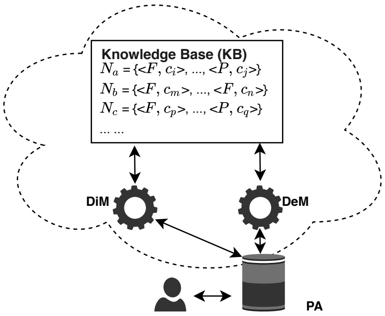
Knowledge Base (KB)
$N_a$ = {<$F$, $c_i$>, ..., <$P$, $c_j$>}
$N_b$ = {<$F$, $c_m$>, ..., <$F$, $c_n$>}
$N_c$ = {<$F$, $c_p$>, ..., <$P$, $c_q$>}
... ...
DiM
DeM
PA

Fig. 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.

## III. Evaluation

We use a fully-anonymized and publicly-available dataset[1] of 292,478 real privacy decisions, which was the result of a survey of PA users in households [3].

We provide details about parameter tuning in [11]. Next, we compare our model with other interpretable and non-interpretable approaches. We also consider a baseline, control condition where the decision is random.

The result comparing the performance of different versions of PEM with previous approaches in the literature can be seen in Table I. 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 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.

TABLE I
Performance comparison with other approaches.

| Model | Interpretable | Accuracy |
|---|---|---|
| $PEM1$: Review every 16 cases | Yes | 0.849 |
| $PEM2$: Review every 64 cases | Yes | 0.840 |
| $PEM3$: No Review | Yes | 0.835 |
| RIVER incremental learning [13] | Yes | 0.772 |
| Zhan et al. [9] | Yes | 0.741 |
| Very fast decision tree (VFDT) [14] | Yes | 0.706 |
| Neural network (MLP) | No | 0.680 |
| Baseline (Random decision) | - | 0.501 |

## References

[1] N. Abdi, K. Ramokapane, and J. 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.

[2] J. Edu, J. Such, and G. Suarez-Tangil, “Smart home personal assistants: a security and privacy review,” *ACM Computing Surveys (CSUR)*, vol. 53, no. 6, pp. 1–36, 2020.

[3] N. Abdi, X. Zhan, K. M. Ramokapane, and J. Such, “Privacy norms for smart home personal assistants,” in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–14.

[4] J. Edu, X. Ferrer-Aran, J. Such, and G. Suarez-Tangil, “Measuring alexa skill privacy practices across three years,” in *Proceedings of the Web Conference (WWW)*, 2022.

[5] Amazon, “Alexa skills,” Video, Mar. 2008. [Online]. Available: https://developer.amazon.com/en-US/alexa/alexa-skills-kit

[6] J. Edu, X. Ferrer-Aran, J. Such, and G. Suarez-Tangil, “Skillvet: Automated traceability analysis of amazon alexa skills,” *IEEE Transactions on Dependable and Secure Computing (TDSC)*, vol. 20, no. 1, pp. 161–175, 2023.

[7] E. Zeng and F. 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)*, 2019, pp. 159–176.

[8] W. Seymour, M. Coté, and J. 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)*, 2023, pp. 166:1–166:16.

[9] X. Zhan, S. Sarkadi, N. Criado, and J. Such, “A model for governing information sharing in smart assistants,” in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 2022, pp. 845–855.

[10] J. Colnago, Y. Feng, T. Palanivel, S. Pearman, M. Ung, A. Acquisti, L. F. Cranor, and N. 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*, 2020, pp. 1–13.

[11] X. Zhan, S. Sarkadi, and J. Such, “Privacy-enhanced personal assistants based on dialogues and case similarity,” in *Proceedings of the European Conference on Artificial Intelligence (ECAI)*, 2023, pp. 670–680.

[12] J. Kolodner, *Case-based reasoning*. Morgan Kaufmann, 2014.

[13] J. Montiel, M. Halford, S. M. Mastelini, G. Bolmier, R. Sourty, R. Vaysse, A. Zouitine, H. M. Gomes, J. Read, T. Abdessalem *et al.*, “River: machine learning for streaming data in python,” 2021.

[14] G. Hulten, L. Spencer, and P. Domingos, “Mining time-changing data streams,” in *Procs of the ACM SIGKDD conference on Knowledge discovery and data mining*, 2001, pp. 97–106.

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