Uncertainty-aware Personal Assistant and Explanation Method for Privacy Decisions

Doctoral Consortium

Gönül Aycı
Bogazici University
Istanbul, Turkey
gonul.ayci@boun.edu.tr

ABSTRACT

In many of today's software systems, most notably online social networks, users can share personal information. Behind the simple action of sharing is a more complicated thought process regarding privacy: which content to share, with whom to share, and why to share. For a user, it's time-consuming and error-prone to check individual personal content for privacy violations. Hence, it would be ideal if a personal assistant can learn its users' privacy preferences and subsequently help users' decision-making by signaling potentially private content. A personalized privacy assistant can help its user make privacy decisions taking into account the ambiguity and uncertainty of privacy predictions as well as its user's personal preferences. Moreover, an explanation of why an image is considered public or private can aid the user in understanding the assistant's decisions.

KEYWORDS

Privacy, Uncertainty, Explainability, Online Social Networks

ACM Reference Format:
Gönül Aycı. 2023. Uncertainty-aware Personal Assistant and Explanation Method for Privacy Decisions: Doctoral Consortium. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 2 pages.

1 INTRODUCTION

Personal assistants have become increasingly important in assisting users in managing their online privacy such as in online social networks (OSNs). This is crucial because users need to have control over the information they share online, and personal assistants can help them achieve that.

Personal assistants for privacy can able to learn users' privacy preferences and detect privacy violations in OSNs by analyzing users' privacy preferences, which are obtained through elicitation [4]. They can determine whether content shared by other users about the user on the network violates the user's privacy preferences. Moreover, when content is owned by multiple users, such as a group image, personal assistants can help users reach privacy decisions by applying negotiation techniques [3, 10]. The use of learning models that extract visual and textual features can also aid in making privacy predictions [9, 11]. Personal assistants can also help its user make privacy decisions by capturing the privacy ambiguity using uncertainty modeling [1]. The personal privacy assistant should recognize when it is uncertain about a decision and instead of making a potentially wrong decision, it should inform the user that it is uncertain and ask the user to make the final decision.

2 PURE: UNCERTAINTY-AWARE PERSONAL PRIVACY ASSISTANT

The personal privacy assistant PURE [1] is designed to help users make informed privacy decisions about their online content by recommending whether it should be labeled private or public. It uses uncertainty modeling to capture situations where privacy is ambiguous and delegate decision-making to the user. It explicitly calculates the uncertainty of its decisions using Evidential Deep Learning (EDL) [7], which quantifies the predictive uncertainty of deep neural networks. PURE is personalized, taking into account each user's risk of misclassification and using their own labeled data to make more accurate privacy decisions. This approach reduces the perceived risk of privacy violations for each user. Additionally, PURE does not require access to any private information of the user or other users in the system, including personal details.

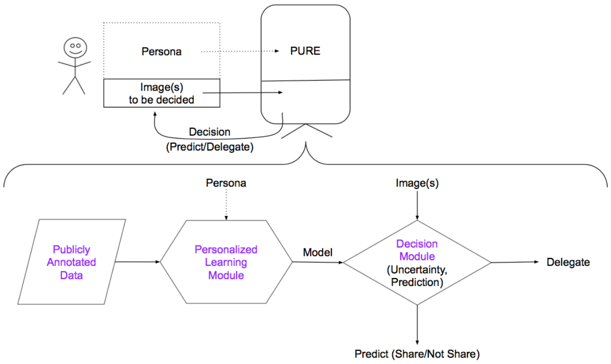


Figure 1: System Overview Schema of PURE[1]

PURE has two modules, the personalized learning module and the decision-making module as shown in Figure 1. The personalized learning module has three objectives: first, it uses publicly labeled data to learn how to classify images as public or private. Secondly, it can detect when it is uncertain about its prediction, and in such cases, delegate the decision to the user. Finally, it takes into account the user's privacy expectations and incorporates them into the learning process, as each user might have different preferences for privacy. This module uses the EDL model to achieve these

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

goals and produces a classification model that can label a given image and estimate the uncertainty in the prediction. Whenever the user provides personally labeled data, this module fine-tunes the personal assistant using the user's images, aiming to decrease the uncertainty that PURE might observe with some images due to the subjective nature of privacy.

The second module of PURE is the decision-making module. When a user needs to make a privacy decision, this module is invoked. This module obtains a prediction and an uncertainty value from the model. The user sets a threshold to decide when to let PURE make a decision or when they want to be involved in the decision-making process. If the uncertainty is above the set threshold, the user decides on the privacy label. Otherwise, the prediction of the model is assigned as the label.

3 EXPLANATIONS FOR PRIVACY PREDICTIONS

Personal privacy assistants help their users to preserve their privacy in OSNs. Personal assistants can also help users better understand the privacy concerns related to their contents (e.g. images) by providing explanations. By doing so, personal assistants can explain how and why a model arrived at its prediction and they can highlight the features of the model that influence a prediction. For example, the saliency map provides a visual explanation by identifying the regions of an input image that are most important in making the prediction [8]. Additionally, the TreeExplainer model provides explanations by showing contributions of features that influence the predictions [5]. Aycı *et al.* [2] propose a methodology to generate explanations as to why a given image is considered public or private. Their explanations are augmented with descriptive text and visually highlight the important features (topics) and related tags (descriptive keywords) in circles. One of their example images is shown in Figure 2.

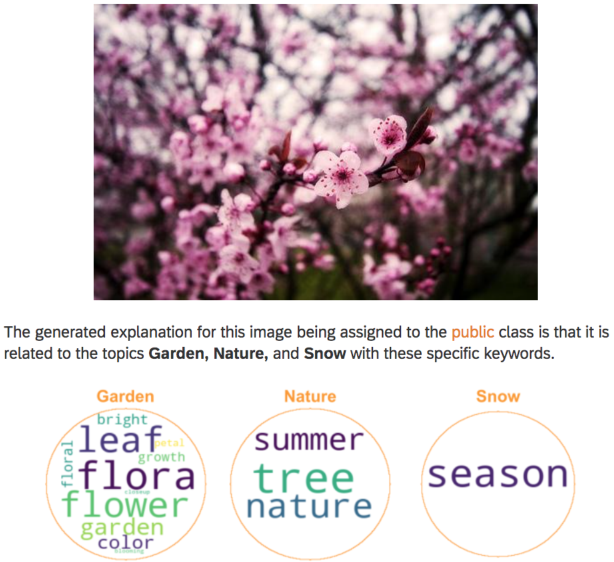
4 FUTURE DIRECTIONS

An important direction is to enable interaction between different personal assistants to help create a collaborative environment for preserving privacy. By doing so, first, a personal assistant can delegate some cases to another assistant to have explanations. Also, this enables making an accurate decision in multi-users scenarios.

Explaining privacy predictions is significant for both end-users and personal assistants. Different factors can have an impact on explanations such as context, location, and so on. However, we do not have access to such external information. Nissenbaum [6] proposes the concept of contextual integrity–that is, people have different expectations of privacy in different contexts, and these expectations can be violated when information is shared in inappropriate ways. An interesting direction is to take into account such factors (if exist) while generating explanations.

Another important direction is to be able to get feedback from users and update the explanations. Privacy is subjective so training the explanation system with the user's own feedback can help to have personalized explanations.

Another interesting direction would be to use personal assistants in different types of content such as confidential documents. This can reduce the complexity of processing such documents. Also, explanations can be generated for different domains such as the medical field.



The generated explanation for this image being assigned to the public class is that it is related to the topics **Garden, Nature,** and **Snow** with these specific keywords.

Figure 2: Example public image and its generated explanation by [2]

REFERENCES

[1] Gönül Aycı, Murat Şensoy, Arzucan Özgür, and Pınar Yolum. 2022. Uncertainty-Aware Personal Assistant for Making Personalized Privacy Decisions. *ACM Transactions on Internet Technology* (August 2022).

[2] Gönül Aycı, Pınar Yolum, Arzucan Özgür, and Murat Şensoy. 2023. Explain to Me: Towards Understanding Privacy Decisions. In *Proceedings of the 22nd Conference on Autonomous Agents and MultiAgent Systems.*

[3] Dilara Kekulluoglu, Nadin Kokciyan, and Pınar Yolum. 2018. Preserving Privacy as Social Responsibility in Online Social Networks. *ACM Transactions on Internet Technology* 18, 4, Article 42 (April 2018), 22 pages.

[4] Nadin Kökciyan and Pınar Yolum. 2016. PriGuard: A Semantic Approach to Detect Privacy Violations in Online Social Networks. *IEEE Transactions on Knowledge and Data Engineering* 28, 10 (2016), 2724–2737.

[5] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* 2, 1 (2020), 2522–5839.

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

[7] Murat Sensoy, Lance Kaplan, and Melih Kandemir. 2018. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems.* 3179–3189.

[8] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).

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

[10] Jose M. Such and Michael Rovatsos. 2016. Privacy Policy Negotiation in Social Media. *ACM Transactions on Autonomous and Adaptive Systems* 11, 1, Article 4 (Feb. 2016), 29 pages.

[11] Ashwini Tonge and Cornelia Caragea. 2020. Image privacy prediction using deep neural networks. *ACM Transactions on the Web (TWEB)* 14, 2 (2020), 1–32.