EDITORS: Pradeep K. Murukannaiah, p.k.murukannaiah@tudelft.nl
Munindar P. Singh, mpsingh@ncsu.edu

DEPARTMENT: INTERNET ETHICS

Can We Explain Privacy?

Gönül Aycı, *Bogazici University, Istanbul, 34342, Turkey*
Arzucan Özgür, *Bogazici University, Istanbul, 34342, Turkey*
Murat Şensoy, *Amazon Alexa AI, EC2A 2FA, London, U.K.*
Pınar Yolum, *Utrecht University, 3584CC, Utrecht, The Netherlands*

Web users want to protect their privacy while sharing content online. This can be done through automated privacy assistants that are capable of taking actions by detecting privacy violations and recommending privacy settings for content that the user intends to share. While these approaches are promising in terms of the accuracy of their privacy decisions, they lack the ability to explain to the end user why certain decisions are being made. In this work, we study how privacy assistants can be enhanced through explanations generated in the context of privacy decisions for the user content. We outline a methodology to create explanations of privacy decisions, discuss core challenges, and show example explanations that are generated by our approach.

Millions of pictures and videos are being shared on social media platforms every day. Our personal data as well as the private content that we create circulate on the Web in ways we haven't imagined before. More and more, cloud services are the go-to locations for storing data, computing, and sharing content. While these services have a lot of benefits for end users, they may use many other third-party services to deliver value, and this may pose unprecedented privacy challenges.

The main method for handling privacy with these services is through consent, where the service provides information on how it will make use of the content, including the purpose and further processing that will be involved and the user of the service is asked to accept the conditions. The informed consent is aimed to detect whether personal data has been leaked or used against the person's will. Current models and implementations of consent prove cumbersome for users. The General Data Privacy Regulation (GDPR)[1] governs privacy and consent in Europe and is influencing other jurisdictions. GDPR requires services to provide explanations, but those explanations are usually long texts. The users are not always clear if their declining to give consent will result in what parts of the service not being available. The services a user engages are numerous and include social media working with documents on the cloud. Thus, many users lack the capacity to even read the text to which they are giving consent.[2]

Privacy assistants can work with humans to help them with tasks related to managing their privacy.[3,4] As users share content, it is necessary to think for whom the content is meant and how to configure its privacy settings.[5] An important category of such content is images. Recent work helps users categorize whether a given image is private or not.[6] This could be useful to help users avoid unintended sharing of private images on social networking sites. Privacy is personal and subjective: What one person identifies as private might be different from that of another. Thus, an assistant ought to provide personalized answers as to whether an image is private or not.[7,8] For these privacy assistants to be adopted by end users, they need to be trusted. One important path to induce such trust is through explanations.[9] We address a novel problem

1089-7801 © 2023 IEEE
Digital Object Identifier 10.1109/MIC.2023.3270768
Date of current version 10 July 2023.

concerning explanations and privacy: How can a privacy assistant explain why it identifies a certain piece of content as private or public?

EXPLANATIONS FOR PRIVACY

Explainable artificial intelligence offers methods that can help humans understand how algorithms work. Most of these methods are targeted to explain how a machine learning algorithm makes a classification. When the classification pertains to an image, visual explanations are useful. An example of a visual explanation would be highlighting the most relevant region in the image (e.g., highlighting a child in an image as to show that there is one). Saliency maps and attention maps are important tools for such explanations.[10] However, these techniques are not immediately applicable to explaining privacy decisions. Specifically, an image is not categorized private or public because of a single segment. For example, an image of a child at home with her parents might be categorized as private, while an image of the same child participating in a school performance on stage might be categorized as public. Hence, identifying and highlighting the child in the image will not adequately explain why this image is private or public.

Another class of explanation techniques works on binary classification and considers what features of the input have been influential on the decision.[11] Thus, they provide a handle to interpret the decision. For example, these methods could say which features had an effect on classifying an image as private and what the strength of these features were. If the features are derived from images automatically, the features might not always translate to concepts that users would understand. Such knowledge may help algorithm developers but is not meaningful for end users. An alternative would be to use image tags for classification. While the tags themselves are understandable for the end user, the number of tags makes it difficult to generate succinct explanations.

METHODOLOGY FOR EXPLAINING PRIVACY

We propose to use the concept of *topics* as a way to capture explanations. We envision that each image belongs to multiple topics with different strengths, where the interplay between the topics leads to understanding why the image is private. For end users, the explanation needs to be either visual or short text and should touch on the most important aspects, rather than giving a comprehensive analysis of features. An explanation as to why an image is private or public will be described through a carefully selected subset of topics that the image belongs to. In order to realize such explanations, we need to understand how we can associate images with topics and how we can decide on which topics to use for explanation.

Our proposed methodology has the following steps:

1) Start with a set of images already labeled as public or private. This could be the set of images that the user herself previously made a decision to share or not to share online.
2) Assign tags (i.e., keywords) for each image to describe its content. These tags could be provided by the users or can be generated automatically with a tool, such as Clarifai.[12]
3) Perform topic modeling. Each topic should pertain to images that could be described with similar tags. At the same time, each topic should be different from each other. Topic modeling is a technique that discovers latent topics within a collection of textual information, in this case, tags associated with images. As a result, each image is associated with one or more topics.
4) Create topic descriptions. Topics are intuitively meaningful and interpretable for humans. In order to improve the understanding for the user, it is useful to name the topics, for example, *Nature*, *Child*, and *Fashion*. This can be done manually as well as automatically. Different automated approaches can be applied such as identifying the most similar word to the tags as the topic name, where similarity can be computed by using the word embeddings of the tags or by using an ontology such as WordNet.[13]
5) Machine learning classification. Using the generated set of topics as features, it is necessary to train an interpretable machine learning model to perform binary classification on the images. Given a new image that is associated with topics, the classifier will assess whether it is private or public.
6) Evaluate the effect of each topic on the classification and determine which topics will be used for explanation. This can be done by different heuristics; for example, by identifying a largely influential single topic or multiple topics that make smaller contributions as well as identifying topics that have opposing classifications.
7) Create textual and visual explanation templates based on the interplay between topics. If only a single topic is influential, it is enough to mention that topic. If, on the other hand, opposing topics are present, the text should express their relation to each other.

Generating Topics

For Step 1, we use a balanced subset of the publicly available PicAlert dataset[14], which is widely used for the privacy prediction for images.[7,15] PicAlert contains Flickr images that are labeled as *private* or *public* by annotators. We consider an image as private if at least one annotator has annotated it as private and public only if all the annotators have annotated it as public. The balanced subset we work with contains 32,000 samples, comprising 27,000 training images and 5000 test images. For Step 2, we use Clarifai application programming interface[12] to automatically generate 20 different tags for each image. For Step 3, we explore 20 different latent topics using *Non-negative matrix factorization*[16] as the topic modeling technique and name the topics based on the dominant keywords (Step 4).

To evaluate the representation of the images with the topics extracted using non-negative matrix factorization (NMF), we train a random forest classifier where the images are represented as term frequency-inverse document frequency (TF-IDF) vectors of these topics (Step 5). We study the precision, recall, and F1 score of the classifier for the private and public class separately. This is important because in many settings the cost of misclassifying a private image might be higher than that of misclassifying a public image.[8] We observe that the scores for both classes are similar. For the private and public class, the classifier obtains a precision of 87% and 89%, recall of 90% and 87%, and F1 score of 89% and 88%, respectively. Overall, the accuracy of the classifier on the test set is 88%, indicating that the NMF-extracted topics are effective for privacy prediction. This performance matches that of state-of-the-art approaches for image privacy prediction[6,8] and thus can be used to generate the explanations.

Generating Explanations From Topics

Even though now we have access to the topics associated with each image and that they are successful in classifying the images, generating explanations from this is still challenging. First, many topics are associated with each image; thus, listing all relevant topics is meaningless. Second, the topics that are mostly associated with the image do not have a clear prediction. Some topics such as *People* are associated more frequently with the private class, whereas others like *Nature* are associated more frequently with the public class. Note that although some topics are associated more frequently with one class, the topics do not have an explicit class to which they belong. Therefore, the topic itself does not directly signal a certain class, and as such it is not straightforward to generate an explanation for the decision by simply looking at its class. For example, the *Performance* topic was associated with 35% of the private images as well as 30% of the public images and the *Competition* topic was associated with 37% of the private images as well as 25% of the public images. Thus, we need to consider to what extent a topic is related to the image as well as how the different topics come together in an image to explain privacy.

The TreeExplainer[17] model provides the contributions of each feature in terms of Shapley values, which affect the model output of tree-based algorithms. A feature with a positive Shapley value denotes that the existence of that feature was influential in the prediction and vice versa for the negative value. For us, each feature corresponds to a topic. We obtained these topic-value pairs from the TreeExplainer. We remove the topics that are not relevant for the image based on its TF-IDF value as well as filter out the topics with values smaller than a threshold (Step 6). The explanations differ on how the remaining topics are related to each other as follows:

Single: It could be that a single topic defines the classification or

Multiple: that multiple topics make small contributions to classification or

Combination: combinations of various topics describe the classification.

For each type, we formulate a short text template and a visual that lists topics and tags that were important for the classification (Step 7).

EXAMPLE EXPLANATIONS

We consider two explanation templates that differ on how the topics relate to each other. The first template pertains to a case where the image has a number of topics, such that none of the topics by themselves would necessarily derive that the image would be of a given target class. However, the existence of multiple topics supports that the image should be of a particular target class. Figure 1 shows an example image that has been identified as private and the explanation generated by our algorithm. Three topics that are relevant to the image are provided with the tags that are important for the classification.

All three topics together strengthen the decision. An image could (and usually does) have many topics associated with it. However, to keep the explanations simple, we select only those topics that contribute the most to the decision. At the same time, it could be possible that the lack of certain topics would affect the underlying decision. For example, the fact that the image is not related to the topic *Outside* might have

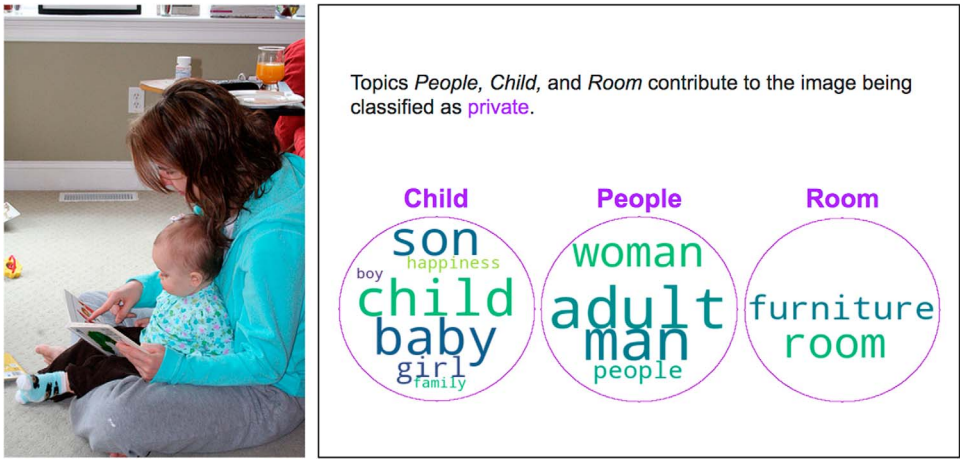

FIGURE 1. Example image classified as private and the generated explanation.

been important in the classification. However, we do not include absent topics as part of the explanation.

The second template pertains to a case where the topics associated with an image do not always agree on whether the image should be private or public. In such situations, the explanation should indicate such opposing evidence as well as how and if it was resolved. One can naively expect that if the image is associated strongly with at least one private topic, then the privacy label of the image would also be private. However, interestingly private concepts when integrated into public space can yield images that are public.

Figure 2 shows an example image that has been identified as public and the generated explanation. Even though the topic *People* pushes this image to be classified as private, the fact that it is situated in the *Art* topic makes it public. By observing that the influence of the *Art* topic is larger than that of *People*, the algorithm can generate the explanation on the right side of Figure 2. Note that the explanation template now is different than that of Figure 1 and reflects that there were opposing topics involved. These examples demonstrate how the interaction between various topics affect the outcome of the classification and thus the explanation that needs to be generated.

DIRECTIONS

The methodology that we propose generates explanations for end users to understand why a given image would be classified as private or public. It is based on exploring hidden topics using topic modeling from descriptive tags of images. It captures explanation templates that are based on the relationship between images and their associated topics and generates explanations automatically. An important direction is to design other explanation templates based on the interplay between topics that push the classification to private or public. Currently, we do not differentiate between topic characteristics; however, some topics, such as *Nature* or *Room*, pertain to the location context, whereas some topics, such as *Competition* or

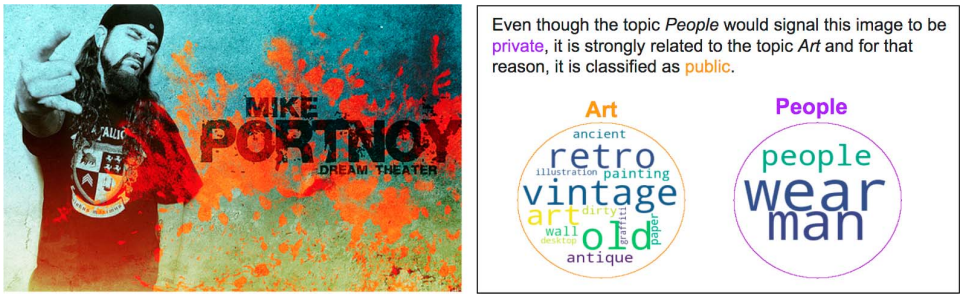

FIGURE 2. Example image classified as public and visual explanation that shows the topics and relevant tags.

Performance denote public spaces. Capturing the semantics of these topics could lead to more detailed and better structured explanations of privacy. Our previous work focused on uncertainty and confidence of predictions,[8] and in this work, we explain model predictions. These two directions for enhancing privacy decisions are complementary and combining them may help the user assess the explanations better. An important direction for future work is to be able to get feedback from end users and update the explanations. With a user study, we plan to understand if the generated explanations make sense to people and if the participants find the explanations useful. This would bring us closer to understand what other aspects need to go into explanations to make them viable for end users.

ACKNOWLEDGMENTS

The first author is supported by the Scientific and Technological Research Council of Turkey (TÜBITAK) and Turkish Directorate of Strategy and Budget under the TAM Project number 2007K12 – 873. This research was partially funded by the Hybrid Intelligence Center, a 10-year program funded by the Dutch Ministry of Education, Culture, and Science through the Netherlands Organization for Scientific Research. This work does not relate to Şensoy's position at Amazon.

REFERENCES

1. C. J. Hoofnagle, B. Van Der Sloot, and F. Z. Borgesius, "The European Union general data protection regulation: What it is and what it means," *Inf. Commun. Technol. Law*, vol. 28, no. 1, pp. 65–98, Jan. 2019, doi: 10.1080/13600834.2019.1573501.
2. T. Vila, R. Greenstadt, and D. Molnar, "Why we can't be bothered to read privacy policies," in *Economics of Information Security*, L. J. Camp and S. Lewis, Eds. Boston, MA, USA: Springer-Verlag, 2004, pp. 143–153.
3. R. L. Fogues, P. K. Murukannaiah, J. M. Such, and M. P. Singh, "SoSharP: Recommending sharing policies in multiuser privacy scenarios," *IEEE Internet Comput.*, vol. 21, no. 6, pp. 28–36, Nov./Dec. 2017, doi: 10.1109/MIC.2017.4180836.
4. J. Colnago et al., "Informing the design of a personalized privacy assistant for the Internet of Things," in *Proc. CHI Conf. Human Factors Comput. Syst.*, 2020, pp. 1–13, doi: 10.1145/3313831.3376389.
5. O. Ulusoy and P. Yolum, "Panola: A personal assistant for supporting users in preserving privacy," *ACM Trans. Internet Technol.*, vol. 22, no. 1, pp. 1–32, Sep. 2021, doi: 10.1145/3471187.
6. A. Tonge and C. Caragea, "Image privacy prediction using deep neural networks," *ACM Trans. Web*, vol. 14, no. 2, pp. 1–32, Apr. 2020, doi: 10.1145/3386082.
7. A. Can Kurtan and P. Yolum, "Assisting humans in privacy management: An agent-based approach," *Auton. Agents Multi-Agent Syst.*, vol. 35, no. 1, pp. 1–33, Apr. 2021.
8. G. Aycı, M. Şensoy, A. Özgür, and P. Yolum, "Uncertainty-aware personal assistant for making personalized privacy decisions," *ACM Trans. Internet Technol.*, vol. 23, no. 1, pp. 1–24, Mar. 2023, doi: 10.1145/3561820.
9. F. Mosca and J. Such, "An explainable assistant for multiuser privacy," *Auton. Agents Multi-Agent Syst.*, vol. 36, no. 1, pp. 1–45, Apr. 2022, doi: 10.1007/s10458-021-09543-5.
10. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 618–626.
11. M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.
12. Clarifai. [Online]. Available: https://clarifai.com/clarifai/main/models/general-image-recognition
13. G. A. Miller, "Wordnet: A lexical database for English," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, Nov. 1995, doi: 10.1145/219717.219748.
14. S. Zerr, S. Siersdorfer, J. Hare, and E. Demidova, "Privacy-aware image classification and search," in *Proc. 35th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2012, pp. 35–44, doi: 10.1145/2348283.2348292.
15. A. Squicciarini, C. Caragea, and R. Balakavi, "Toward automated online photo privacy," *ACM Trans. Web*, vol. 11, no. 1, pp. 1–29, Apr. 2017, doi: 10.1145/2983644.
16. D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, Oct. 1999, doi: 10.1038/44565.
17. S. M. Lundberg et al., "From local explanations to global understanding with explainable AI for trees," *Nature Mach. Intell.*, vol. 2, no. 1, pp. 56–67, Jan. 2020, doi: 10.1038/s42256-019-0138-9.

GÖNÜL AYCI is a Ph.D. student at the Computer Engineering Department, Bogazici University, Istanbul, 34342, Turkey, and a visiting Ph.D. Researcher at Utrecht University. Her

research interests include designing and developing privacy assistants. Ayci received her M.Sc. degree in computer science from Ozyegin University, Istanbul, Turkey. Contact her at gonul.ayci@boun.edu.tr.

ARZUCAN ÖZGÜR is a faculty member at the Computer Engineering Department, Bogazici University, Istanbul, 34342, Turkey. Her research interests include natural language processing and bioinformatics. Özgür received her Ph.D. degree in computer science and engineering from the University of Michigan. Contact her at arzucan.ozgur@boun.edu.tr.

MURAT ŞENSOY is an applied research scientist at Amazon Alexa AI, EC2A 2FA, London, U.K. His research interests include reliable machine learning systems. Şensoy received his Ph.D. degree in computer engineering from Bogazici University. Contact him at drmuratsensoy@gmail.com.

PINAR YOLUM is a professor in information and computing sciences at Utrecht University, 3584CC, Utrecht, The Netherlands. Her current research interests include trustworthy AI with an emphasis on privacy. Yolum received her Ph.D. degree in computer science from North Carolina State University. Contact her at p.yolum@uu.nl.