TOWARDS TRUSTWORTHY PERSONAL ASSISTANTS FOR PRIVACY

by

Gönül Aycı

B.S., Mathematics, Marmara University, 2013

M.S., Computer Science, Ozyegin University, 2016

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

Graduate Program in Computer Engineering

Boğaziçi University

2023

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my advisors, Assoc. Prof. Arzucan Özgür and Prof. Pınar Yolum. Their pleasant way of communicating, insightful guidance, and constructive feedback throughout the entire research journey have been invaluable. They have helped me improve my skills tremendously but they have also helped me grow on a personal level to shape the person I am today. And for that, I am very grateful as well. I would like to express my sincere gratitude to the thesis committee Prof. Lale Akarun, Assist. Prof. Fatma Başak Aydemir, Assoc. Prof. Reyhan Aydoğan, and Prof. Şule Gündüz Öğüdücü for the time and commitment in shaping this work.

I would like to express my heartfelt thanks to my partner, Cas; my father, Mehmet; my mother, Menekşe; my sister, Gamze; my dear brother-in-law Orhan and my brothers, Gökhan, Gürkan, and Güray. As the first in my family to attend university and pursue a doctorate, I am truly indebted to my family for their endless support and encouragement throughout this journey.

I want to give special thanks to Assoc. Prof. Murat Şensoy for the great collaboration. The things that I have learned from him helped me broaden my horizons. Also, I would like to thank Prof. Taylan Cemgil and Dr. Suzan Üsküdarlı for the opportunity to work on the Horizon 2020 European Union *OpenMaker* project together. I gained invaluable practical experience and knowledge.

Special thanks to my colleagues in the Computer Science departments of Bogazici University and Utrecht University for the many tea breaks, and for sharing their knowledge with me. And a very special thanks to my close friends that make me the happiest person on Earth with so much joy and happiness. I feel very fortunate to have such great colleagues and friends.

I would like to thank the Scientific and Technological Research Council of Turkey (TÜBİTAK) for the visiting scholarship to Utrecht University in the Netherlands. Secondly, I would also like to thank the Turkish Directorate of Strategy and Budget under the TAM Project number $2007K12-873$ for supporting my work.

ABSTRACT

TOWARDS TRUSTWORTHY PERSONAL ASSISTANTS FOR PRIVACY

Many software systems, such as online social networks, enable their users to share information about themselves online. However, users worry about the privacy implications of sharing content. It's a tedious process to make privacy decisions and it makes managing privacy difficult. Recent approaches to help users manage their privacy involve building personal assistants that can recommend whether a user's content is private or not. However, privacy's ambiguous nature and difficulties in explaining assistants' decision-making are challenges hampering users' trust in these systems and therefore also widespread user adoption. In this dissertation we design trustworthy privacy assistants that can help tackle both challenges. We first propose a personal assistant called PURE that integrates machine learning to make predictions on whether a user would identify an image as private or not. An important characteristic of PURE is its ability to model uncertainty in its decisions explicitly. When uncertainty is high, no prediction is made and the decision is delegated to the user. By factoring in user's own understanding of privacy, PURE is able to personalize its recommendations.

A second crucial factor in fostering trust in personal assistants is their ability to explain their decision-making processes. Our second assistant PEAK is capable of generating such explanations for its recommendations, using latent topics and predefined explanation categories to do so. A user study shows users find PEAK's explanations useful and easy to understand. Additionally, privacy assistants can use the explanations to improve their own decision-making, with the incorporation of PEAK into PURE resulting in less uncertain images delegated to the user whilst model performance is not compromised. Overall, our work makes an important contribution towards the development of trustworthy personal assistants capable of preserving users' privacy.

ÖZET

MAHREMİYET İÇİN GÜVENİLİR KİŞİSEL ASİSTANLARA DOĞRU

Çevrimiçi sosyal ağlar gibi birçok yazılım sistemi, kullanıcılarının çevrimiçi olarak kendileri hakkında bilgi paylaşmasına olanak tanır. Ancak kullanıcılar, içerik paylaşmanın mahremiyetle ilgili sonuçlarından endişe duyarlar. Bu kararları vermek zahmetli bir süreçtir ve mahremiyet yönetimini zorlaştırır. Kullanıcıların mahremiyetlerini yönetmelerine yardımcı olan son yaklaşımlar, bir kullanıcının içeriğinin mahrem olup olmadığını önerebilen kişisel asistanlar inşa etmeyi kapsar. Ancak mahremiyetin muğlak doğası ve asistanların karar verme sürecini açıklamadaki güçlükler, kullanıcıların bu sistemlere olan güvenini ve bu nedenle yaygın bir şekilde benimsenmesini engelleyen zorluklardır. Bu tezde, her iki zorluğun da üstesinden gelmeye yardımcı olabilen güvenilir mahremiyet asistanları tasarlıyoruz. İlk olarak, kullanıcının mahremiyet kararları almasına yardımcı olan PURE isminde bir kişisel asistan öneriyoruz. PURE'ün önemli bir özelliği, kararlarındaki belirsizliği açıkça modelleme yeteneğidir. Belirsizlik yüksek olduğunda, tahmin yapılmaz ve karar kullanıcıya bırakılır. PURE, kullanıcının mahremiyet anlayışını dikkate alarak önerilerini kişiselleştirebilir.

Kişisel asistanlara duyulan güveni artırmada ikinci önemli faktör, karar verme süreçlerini açıklama yetenekleridir. İkinci mahremiyet asistanımız PEAK, gizli konuları ve tanımlanan açıklama kategorilerini kullanarak önerileri için açıklamalar üretebilmektedir. Bir kullanıcı çalışması, kullanıcıların PEAK'in açıklamalarını yararlı ve kolay anlaşılır bulduğunu göstermektedir. Ayrıca, mahremiyet asistanları, açıklamaları kendi karar süreçlerini iyileştirmek için kullanabilir; bu, PEAK'i PURE'e entegre ederek gösterilmiştir ve bu durumda belirsizlik içeren resimler kullanıcıya devredilirken model performansı etkilenmez. Genel olarak, çalışmamız, kullanıcıların mahremiyetini koruyabilen güvenilir kişisel asistanların geliştirilmesine önemli bir katkı sağlamaktadır.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS iii
ABSTRACT v
ÖZET vi
LIST OF FIGURES ix
LIST OF TABLES xii
LIST OF SYMBOLS xiv
LIST OF ACRONYMS/ABBREVIATIONS xv
1. INTRODUCTION 1
1.1. Research Objectives 3
1.2. Contributions and Thesis Outline 9
2. UNCERTAINTY-AWARE PERSONAL PRIVACY ASSISTANT 12
2.1. Preserving Privacy with PURE 14
2.1.1. Learning Privacy Labels with Uncertainty 15
2.1.2. Personalizing Privacy 20
2.2. Experimental Setup 24
2.3. Results 27
2.4. Conclusion 41
3. IMAGE PRIVACY PREDICTION USING TEXTUAL FEATURES 42
3.1. Understanding Underlying Factors 43
3.1.1. Tag Identification 46
3.1.2. Tag Generation 48
3.2. Extracting Topics 50
3.3. Evaluation 51
3.4. Conclusion 54
4. EXPLAINABLE PRIVACY ASSISTANT 55
4.1. Preserving Privacy with PEAK 57
4.2. Generating Explanations from Topics 60

4.2.1. Computing the Contribution of Topics using the TreeExplainer Model 60
4.2.2. Identifying Explanation Categories of Images 62
4.2.3. Computing Explanations 66
4.3. Evaluation 71
4.4. Conclusion 81
5. LITERATURE SURVEY 83
5.1. Uncertainty and Risk Models 85
5.2. Privacy Prediction Approaches 87
5.2.1. Tag-based Approaches 87
5.2.2. Textual- and Visual-based (Hybrid) Approaches 87
5.2.3. Policy-based Approaches 88
5.2.4. Taxonomy-based Approaches 89
5.2.5. Group-based Approaches 90
5.3. Explanation Models 91
5.3.1. Model-specific Methods 93
5.3.2. Model-agnostic Methods 93
5.3.3. Example-based Explanations 94
6. DISCUSSION 96
6.1. Summary of Contributions 96
6.2. Threats to Validity 97
6.3. Future Directions 98
REFERENCES 101
APPENDIX A: On the Figures of This Dissertation 116

LIST OF FIGURES

Figure 1.1. Examples of images labeled as private and public by the annotators. 5

Figure 1.2. Examples of images to make privacy predictions. 6

Figure 2.1. System overview of PURE. 14

Figure 2.2. An example for predicting privacy labels of four images and quantifying uncertainty values for each prediction. 20

Figure 2.3. An example for predicting privacy labels of four images for a sensitive user and uncertainty values for each prediction. 22

Figure 2.4. An example for predicting privacy labels by fine-tuning using personal data and uncertainty values for each prediction. 23

Figure 2.5. Uncertainty distributions for the private and public classes. 32

Figure 2.6. The change of Accuracy with respect to the uncertainty threshold. 33

Figure 2.7. The change of Accuracy for different models with respect to different entropy thresholds. 35

Figure 2.8. F1 scores for the private and public classes relative to the percentage of delegated decisions. 36

Figure 2.9. The delegation rate by round for different uncertainty thresholds. . 40

Figure 3.1. Example image and its generated tags by Clarifai. 43

Figure 3.2. An example of images and their tags. 44

Figure 3.3. r matrix for Example 3.1. 45

Figure 3.4. s matrix for Example 3.1. 46

Figure 3.5. *tfrf* matrix for Example 3.1. 46

Figure 3.6. Non-negative Matrix Factorization concept. 50

Figure 3.7. Tag clouds for Topics Nature, Child, Performance, Business, and Fashion. 53

Figure 3.8. Percentage of occurrence of each topic in private and public images. 53

Figure 4.1. Workflow of PEAK. 57

Figure 4.2. An example image annotated as private and its generated explanation by PEAK. 59

Figure 4.3. Example output of the TreeExplainer model. 61

Figure 4.4. Example image annotated as public and its generated explanation with the topics Design and Child (Opposing category). 64

Figure 4.5. Example image annotated as private and its generated explanation with the topics People, Fashion, and Room (Collaborative category). 65

Figure 4.6. Example image annotated as private and its generated explanation with the topics People, Business, and Seaside (Weak category). . . 66

Figure 4.7. The answers for all questions. 73

Figure 4.8. The answers for the questions with respect to the public class. . . 74

Figure 4.9. The answers for the questions with respect to the private class. . . 75

Figure 4.10. The answers for the questions with respect to the categories. . . . 76

Figure 4.11. System overview schema of PEAK in combination with PURE. . 78

LIST OF TABLES

Table 2.1. Performance of PURE using different pre-trained models ResNet50, InceptionV3, and VGG16. 27

Table 2.2. Overall results for PURE as training samples are reduced. 28

Table 2.3. Results for the private and public classes of PURE at different training sample rates. 29

Table 2.4. Overall results for PURE at prediction delegation rates 0%, 10%, 25%, 50%, and 75% based on uncertainty. 29

Table 2.5. Results for the private and public classes of PURE at various prediction delegation rates (0%, 10%, 25%, 50%, and 75%) based on uncertainty. 30

Table 2.6. Results for different risk personas. 37

Table 2.7. Ratios of predictions whose uncertainty values are greater than 0.7 (θ). 39

Table 3.1. Comparison of privacy prediction performance of the classifier by using different tag generation methods such as VGG, ResNet, GoogleNet, AlexNet, and Clarifai. 48

Table 3.2. Comparing privacy prediction performance obtained through various input preparation methods, including the TF–RF, as well as word embedding techniques like BERT, GloVe, and Word2Vec. . . 49

Table 4.1. Percentage (%) of all and uncertain images in the Train-set that belong to the explanation categories. 79

Table 4.2. Explanation category and class-specific privacy prediction performance (% accuracy) of PEAK on all and uncertain images in the Train-set. 80

LIST OF SYMBOLS

α	A positive parameter of the Beta distribution
β	A negative parameter of the Beta distribution
δ	Cosine Similarity threshold
θ	Uncertainty threshold
log	Logarithm function

LIST OF ACRONYMS/ABBREVIATIONS

BERT	Bidirectional Encoder Representations from Transformer
CNN	Convolutional Neural Network
DNN	Deep Neural Network
EDL	Evidential Deep Learning
GloVe	Global Vectors
IoT	Internet of Things
KL	Kullback-Leibler
LDA	Latent Dirichlet Allocation
ML	Machine Learning
NMF	Non-negative Matrix Factorization
OSN	Online Social Network
PCA	Principal Component Analysis
ReLU	Rectified Linear Unit
ResNet	Residual Network
RL	Reinforcement Learning
RNN	Recurrent Neural Network
SIFT	Scale-Invariant Feature Transformation
SL	Subjective Logic
SVM	Support Vector Machine
TF–IDF	Term Frequency–Inverse Document Frequency
TF–RF	Term Frequency–Relevance Frequency

1. INTRODUCTION

Privacy has been of profound importance for humanity throughout history. One significant milestone during the early ages of the computer era is the concept of information privacy, which is defined by Westin [1] as "the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others". Westin's investigation of privacy focused on individual freedom and personal autonomy. Warren and Brandeis [2] emphasized that "the right to life has come to mean the right to enjoy life, —the right to be let alone; the right to liberty secures the exercise of extensive civil privileges". Similarly, Douglas stressed the connection between personal liberty and the right to privacy, stating that "the right to be let alone is indeed the beginning of all freedom" [3]. The existence of the "right" to privacy poses a dilemma regarding the level of individuals' engagement in society. O'Brien explored privacy from the perspective of legal and public policy considerations, highlighting it as "a state of limited access to a person" [4]. Beyond these definitions, the notion of privacy and its implications have gained significant importance in recent years. Privacy evolves as an essential facilitator for trust and freedom in today's information society [5].

During the 21st century, many of the computing systems are facilitating and encouraging the sharing of information on a large scale. Billions of users on online social networks (OSNs) are exchanging images, videos, and comments with each other constantly. At the same time, we rely on the Internet of Things (IoT) applications, such as smart home systems, for many everyday activities, which in return collect, store, and fuse our everyday data to provide their services. On one hand, these systems have become undeniably beneficial, but on the other hand, people are finding it challenging to manage their privacy while using them [6].

Several studies suggest that before the explosive growth of the Internet, individuals have always had a desire for privacy [7, 8]. However, the widespread use of

systems such as OSNs and IoT applications has become pervasive due to the provision of useful services, such as document sharing and home entertainment, but users are increasingly concerned about their privacy and often self-censor or delete content after sharing [9]. Privacy concerns such as stalking, identity theft, digital dossier aggregation, unauthorized data access, the misuse of personal data, difficulty of complete account deletions, and linkability from image metadata have become prevalent [10, 11]. For instance, the ability to tag images with metadata on OSNs allows users to include personal information like names and profile links, posing a potential privacy risk whilst the lack of privacy controls for image tagging and accessibility in search results adds to this concern [10]. Moreover, privacy breaches can arise not only from external origins but also without any malicious intent behind it. For instance, OSN users' own privacy settings can be incorrect. As a result, privacy can be at risk, potentially even without the users knowing their settings are incorrectly configured. Even when users are aware of these incorrect settings, a significant majority of them either lack the ability or choose not to fix such errors because the error is for example not deemed to be worth correcting or is not that serious [12]. Indeed, some users do not trust themselves to manage their privacy configurations or do not understand how it would affect their online presence [13]. Another important point relates to users often not having the time or energy to make informed decisions for their vast amount of content, which results in decision fatigue and makes users' decisions error-prone. Lastly, these difficulties in managing privacy are compounded by the fact users must constantly decide whether they want to share content or not, and the discrepancy between users' privacy preferences and their actual behavior [14, 15]. This latter dichotomy is often referred to as the "privacy paradox" and describes how people generally claim to be very concerned about their privacy but they are not behaving as such and are in fact doing little to protect their data and privacy [16, 17]. Privacy assistants can potentially help on both fronts. On the one hand, they can provide recommendations to users, thus saving them time and energy. On the other hand, users' privacy preferences can potentially be incorporated into the assistants' recommendations, possibly helping to bridge the gap between users' stated preferences and actual behaviour.

We understand privacy management as the task of handling who has access to personal information, for what purposes, and in which context. It includes protecting users from being adversely affected by information shared about them, for example, information shared without their consent or awareness on OSNs. Interestingly, the information that can violate a user's privacy can come from different sources and in different forms. It could be that users post information about themselves that later causes them harm, such as losing their job or being investigated due to controversial comments [18,19]. Additionally, others might post content about the user on the user's profile page without their knowledge or consent, such as tagging identifiable, geotagged images [20].

1.1. Research Objectives

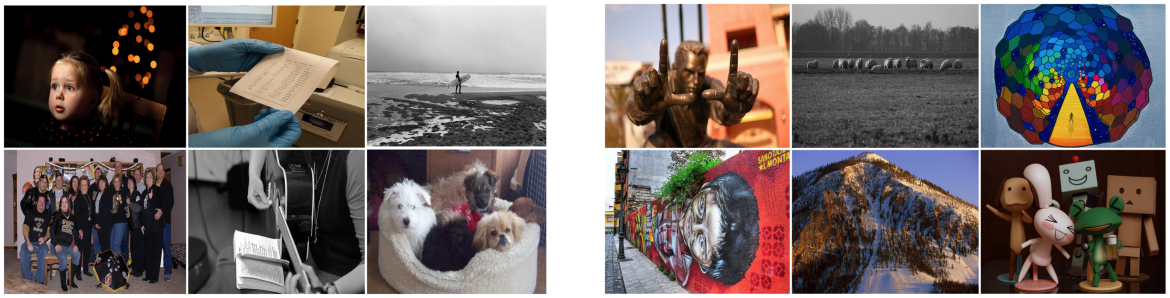
Millions of images and videos are being shared on social media platforms every day. Our personal data as well as the private content that we create circulate on the Web in ways we haven't imagined before. More and more, cloud services are the go-to locations for storing data, computing, and sharing content. 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 its purpose, and further processing that will be involved. The user of the service is then asked to accept the conditions. Current models and implementations of consent prove cumbersome for users. The General Data Privacy Regulation (GDPR) [21] governs privacy and consent in Europe and is influencing other jurisdictions. GDPR requires services to provide explanations of their privacy policies but those explanations are usually long, complicated texts. It's not always clear for the user what the consequences are of declining to give consent, for instance which parts of the service will subsequently be unavailable. Users typically use numerous services, including those working with documents on the cloud. Thus, many users lack the capacity to even read the text to

which they are giving consent [22].

There are further challenges for users, even when they understand the privacy policies they consented to. Thinking about these policies as well as the privacy implications for each piece of content to be shared is tedious. Privacy assistants can work with humans to help them manage their privacy [23, 24]. Personal privacy assistants have been used to help users in various tasks, including time management [25], smart homes [26], voice-assistance [27], and so on. Recently, personal assistants have been used for helping users manage their privacy online. Kökciyan and Yolum [20] develop personal assistants to detect privacy violations in OSNs on behalf of their users. They assume that the personal assistant has access to users' privacy preferences through elicitation. Using these preferences and a domain ontology, the personal assistant computes whether others in the OSN share content about the user against her preferences. Kekulluoglu *et al.* [28] and Such and Rovatsos [29] develop techniques to help users reach privacy decisions when a content being shared is owned by multiple users, such as a group image. Both approaches assume that the personal assistants of the users know the privacy preferences of the users and then they apply negotiation techniques to enable the personal assistants to reach a sharing decision that both users are comfortable with. Ideally, a personal assistant should be able to adjust its recommendation based on a given user's preferences and their personal understanding of privacy. However, even then, there may be situations and content for which it's difficult for the assistant to make a privacy recommendation. Hence, it's important the personal assistant is able to assess when its recommendation is uncertain and subsequently let the user decide. By doing so, it builds the user's trust in the personal privacy assistant, which is a crucial component for eventual adoption.

Nowadays, an important category of content is images, and when sharing an image, users have to think about the intended audience and how to configure the privacy settings appropriately [30]. The image privacy prediction task aims to make a decision about whether a given image is private or public. Figure 1.1 shows examples of images annotated as private and public in the PicAlert dataset [31] by different



(a) Private images (b) Public images

Figure 1.1. Examples of images labeled as private and public by the annotators.

annotators. Recent work helps users categorize whether a given image is private or not [32]. This could be useful to help users avoid unintended sharing of private images on social networking sites. However, determining whether an image is private or not is not that easy. Privacy is very subjective and what one person considers to be private may not be private for someone else. This ambiguity of privacy can be best illustrated using an example. Figure 1.2 shows three example images from the PicAlert dataset. For the first image, the annotators identify this as a private image. In contrast, the annotators identify the second image containing a frozen bike on a mountain range as a public image. However, for the third image, the task of determining the image's privacy label is more difficult. On the one hand, the image looks like it depicts a concert, which suggests we can publicly share it. On the other hand, a user might prefer to keep the image private in order to protect people's identities. It is difficult to reach a consensus on the privacy label of the third image. The fact there are people in the image does not necessarily make it private as there are plenty of examples of public settings that include people. Hence, context matters, which makes the image classification problem more challenging than just an object detection problem. Moreover, given the subjectivity of privacy, the same image might be classified differently depending on the individual (e.g. OSN user).

Ultimately, the objective is to model the ambiguity of privacy predictions. Hence, a privacy assistant should be able to understand a user's personal perception of privacy, and thus ought to provide personalized answers as to whether an image is private or not [33]. Additionally, the ambiguity of privacy means there is inherent uncertainty in



Figure 1.2. Examples of images to make privacy predictions.

attempting to classify the privacy label of images. Hence, a personal assistant should also be able to recognize when there is too much uncertainty surrounding a particular image's privacy label and subsequently refrain from making a privacy prediction. A personalized and uncertainty-aware personal assistant should help to foster trust with users, ultimately leading to the adoption of such privacy assistants. However, another important path to induce trust is through explanations [34]. Indeed, given the ambiguity of privacy, we ideally want to know why the assistant classifies a given image as public or private. If users are able to understand the decision-making, trust in the system can increase.

The importance of being able to understand an assistant's underlying decision-making process and how it relates to trust can be best illustrated with a famous example. Ribeiro *et al.* [35] illustrate how explanations can lead to important insights using an experiment where they train a neural network to distinguish between wolves and huskies. They deliberately trained a "bad" classifier using images of wolves with snow in the background and images of huskies without any snow. As a result, the model predicted *wolf* if there was snow in the image, and *husky* otherwise, regardless of for instance the position or colour of the animal. First of all, the explanation clearly highlights an error in the classifier's decision-making process, which might otherwise have been more difficult to detect. Second of all, and most importantly, they asked a number of subjects whether they trusted the model. When no explanation was shown,

10 out of 27 subjects trusted the "bad" model, but when explanations were shown only 3 out of 27 subjects trusted the classifier. Hence, explanations allowed users to better gauge whether to trust the model or not and why to trust or distrust it. Overall, the experiment illustrates the importance of explanations in trustworthiness.

Explainable Artificial Intelligence (XAI) suggests approaches that aid in the comprehension of why and how a machine learning (ML) model arrives at its prediction. There are various explanation methods with respect to visual and textual explanations [36–39]. For the image privacy prediction task, an example of a visual explanation can be highlighting the most important region in the image for the target class, whereas a textual explanation can be generated text such as "if a guitar had not been in the image, the image would not have been public" [39]. A visual explanation can quickly point out a potential privacy concern, however, it can be overwhelming for users to understand underlying mechanisms since it may present many intricate details [36, 37, 40–42]. Szymanski *et al.* [43] suggest that the combination of visual and textual explanations improves understanding for non-expert (without any software or domain knowledge) users.

Personal assistants can play a pivotal role in helping users while making decisions to ease their online interactions. Although, as mentioned above, recent studies have shown promising progress in tackling tasks such as privacy violation detection and personalized privacy recommendations, a crucial aspect for widespread user adoption is the capability of these systems to be uncertainty-aware and explainable. In this dissertation, we address these challenges in the context of our main research objective of designing trustworthy privacy assistants to manage privacy.

In order to achieve our research objective, we have identified two research questions:

Research Question 1: How can we design an uncertainty-aware privacy assistant to help users make privacy decisions?

Existing personal privacy assistants that learn users' privacy preferences do not incorporate uncertainty in their predictions [32, 33, 44]. However, a personal privacy assistant may face challenges in determining, on behalf of the user, whether content is private or not. In such cases, it is crucial for the assistant to possess an awareness of its own uncertainty and refrain from making uncertain decisions. Ideally, images with uncertain privacy predictions should be delegated to the user for a decision whilst images with (more) certain predictions should be predicted by the personal assistant. As a result, if we consider only certain images, we would expect to obtain a higher accuracy than the overall accuracy for all images. Moreover, for effective personalization, a personal assistant should consider two crucial aspects. First, it should be capable of discerning the privacy expectations of its own user. This factor is important since privacy is subjective, where what is considered private by one user may be considered public by another user. Second, each user possesses a distinct level of risk tolerance when it comes to making erroneous decisions. In this context, "risk" refers to the potential misclassification of content as either public or private. For instance, some users may prefer a risk-averse personal assistant that leans towards classifying content as private, even if there is a chance that the user would prefer it to be public.

Research Question 2: How can we design a privacy assistant that generates an explanation as to why content is considered public or private?

While being aware of prediction uncertainty is a necessary property of privacy assistants, this alone is not enough. In order to cultivate trust and ultimately enable adoption by end users, it is important to design *explainable* privacy assistants. The utilization of semantic information (e.g. descriptive keywords or meta topics) plays a vital role in the advancement of explainable privacy assistants that provide explanations as to why certain decisions are being made. While the textual features of content, e.g. descriptive keywords, themselves are understandable for the end-user, the number of features makes it difficult to generate succinct explanations. Grouping keywords based on semantic similarity can reduce the feature space that enable us to learn privacy preferences without compromising the prediction performance. Then, we should

conduct a user study to assess participants' perception of the usefulness of the generated explanations, as well as to identify the aspects of the explanation and image that influence users' understanding of the model's decision. Moreover, explainable privacy assistants can enhance the performance of privacy assistants' decision-making by working together. Instead of delegating the decision for uncertain images back to the user, an assistant can first delegate the uncertain images to an explainable assistant for privacy predictions.

1.2. Contributions and Thesis Outline

In light of these research questions, we propose two novel privacy assistants to preserve privacy of users engaging in content sharing activities. In order to design our assistants, we focus on image sharing; as widely done in OSNs. First, we present an uncertainty-aware personal assistant called PURE that explicitly quantifies uncertainty in its predictions, and when the prediction is highly uncertain, delegates the decision-making back to its user. Second, we present an explainable privacy assistant called PEAK that generates explanations for why an image is considered public or private. By incorporating uncertainty-awareness and the ability to explain predictions, we are able to design trustworthy privacy assistants capable of helping users manage their privacy.

The main contributions of this dissertation are listed below, following the structure of the thesis:

- In Chapter 2, we tackle the first research question and introduce an uncertainty-aware personal privacy assistant PURE that helps its user to make privacy decisions. While PURE makes a prediction (i.e. *private* or *public*) for each image, it also captures the ambiguity of privacy by calculating a level of uncertainty for that prediction. PURE has access to a collection of data that has been labeled by various annotators (i.e. PicAlert dataset). Using only the visual characteristics of images, PURE learns users' privacy preferences and creates a model. To

make a privacy decision, PURE compares the uncertainty level to a threshold level provided by the user, and if the threshold is exceeded, PURE delegates the privacy decision to the user. Otherwise, PURE uses its own prediction results (i.e. share or not share). Moreover, it takes into account the user's "persona" encompassing their risk perception, personally labeled data, and preferences for consultation. This personalized approach ensures that PURE adapts its behavior for each user, minimizing their perceived risk of privacy violations. Another crucial characteristic of PURE is that it does not require access to any additional private user information, such as personal details.

- Chapter 3 examines how semantic information can be used for making privacy predictions. Utilizing semantic information provides additional insights and allows to build an explainable system. First, we extract the underlying factors on making privacy predictions. We examine the importance of word-embedding methods by leveraging textual features (i.e. tags) of images. Then, we group these factors (i.e. reduce feature space) and discover latent meta (high-level) contextualized features based on their semantic similarity. It enables generating comprehensive and succinct explanations on how decision-making algorithms work. In order to do so, we present a method that extracts meta features (topics) from image-tag relation by topic modelling.
- In Chapter 4, we tackle the second research question and propose an explainable privacy assistant PEAK. We address a novel problem concerning explanations and privacy: How can a privacy assistant explain why it identifies a certain piece of content as public or private? PEAK employs topic modeling techniques while extracting latent topics 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. In order to evaluate PEAK, we first conduct a user study to show that the generated explanations by PEAK sufficient, satisfying, and understandable for humans. Then, we also show that the generated explanations of PEAK are used by personal assistants to improve their decision making. PEAK is able to reduce the number of delegated images to the user without compromising the accuracy of privacy decisions.

- Chapter 5 provides a literature review on image privacy prediction approaches, uncertainty, and risk methods, and explanation models, with comparisons to our work.
- Chapter 6 draws the discussion of the thesis, addresses potential threats to the validity of our study, and states future directions.

2. UNCERTAINTY-AWARE PERSONAL PRIVACY ASSISTANT

There is a tremendous need to learn users' privacy preferences accurately since many approaches depend on using users' privacy preferences. If a personal privacy assistant can represent the privacy expectations well, then these privacy assistants can help the users in privacy dealings, such as warning the user when the user attempts to share a private image, negotiate with other users on behalf of the user, and so on.

While learning the privacy preferences of a user resembles a classical machine learning problem, there are two properties of privacy that make the problem difficult. First, privacy by definition is ambiguous, making it challenging to specify. This makes the pattern that is searched malleable. Second, the users themselves are not always certain about their own privacy preferences and may change their preferences based on other motives [45]. For these reasons, using a traditional predictive model is unreliable as the cost of making a wrong privacy decision is high.

Ideally, the personal privacy assistant should adhere to the following properties:

- Unobtrusive: The privacy assistant should learn from the sharing behavior of the user without interrupting the user (e.g. asking the user what to share or not) as well as without requiring additional information about the user or the image, such as age or occupation of the user or the tags of an image. Thus, the privacy assistant should only consult the user if necessary.
- Uncertainty-aware: As mentioned above, privacy decisions are many times ambiguous. A personal privacy assistant may not always be able to decide if an image is private or not for the user. The assistant should be aware of this uncertainty and be able to say "I don't know" rather than making an uncertain decision. Hence, it should let the user know that it is uncertain and delegate the decision back to the user.

- Personalized: There are two aspects of personalization that are important for the personal assistant to consider. First, the assistant should be able to understand the privacy expectations of its own user. This is important because privacy is subjective and what one user considers private might not be private for another user. Second, each user has a different *risk* associated with making a wrong decision. The risk here refers to classifying a image as private when it should have been public and vice versa. For example, a user might prefer a personal assistant to be risk-averse and classify image as private when there is a chance that it's public.

Existing privacy personal assistants that learn users' privacy preferences do not address the uncertainty of their predictions while making decisions [32, 33, 44]. The idea of considering the risk to personalize decisions has been used before but has not been coupled with privacy decisions as we have done here [46]. Accordingly, this chapter proposes a personal privacy assistant (PURE) that helps its user to make privacy decisions in a personalized way, taking into account the ambiguity of privacy predictions. An important aspect of PURE is that it explicitly calculates the uncertainty of its predictions. When PURE is uncertain of its decisions, it delegates the prediction back to the user. PURE uses publicly annotated data to create an initial model but also a person's understanding of risk, personally labeled data, and when she should be consulted. In this way, PURE behaves differently for each user to minimize the user's perceived risk of privacy violations. Moreover, PURE does not need to have access to any other private information of the user (e.g. personal details or usage patterns) as well as any of the users in the system, including the relations among users.

Organization. Section 2.1 explains our approach in detail. Section 2.2 provides details on the evaluation setup. Section 2.3 evaluates the proposed approach on a widely used data set and demonstrates the benefits of capturing and exploiting uncertainty. Finally, Section 2.4 concludes our work with pointers to future directions.

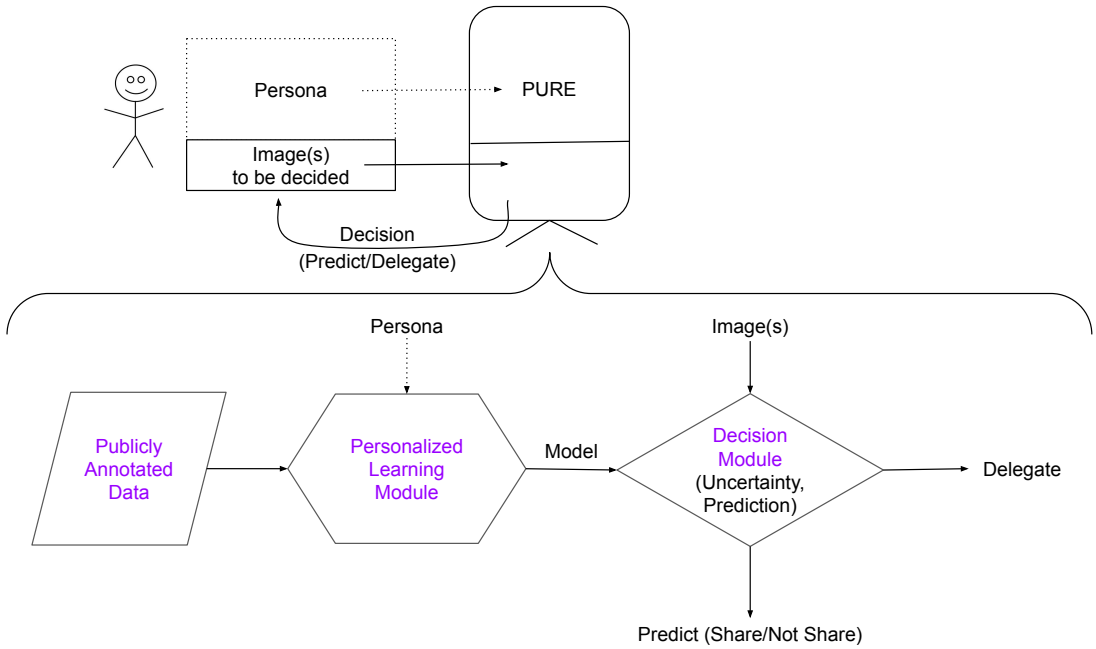


Figure 2.1. System overview of PURE.

2.1. Preserving Privacy with PURE

We envision PURE to work side-by-side with its user when the user is about to share image and help its user make privacy decisions. Figure 2.1 shows a system overview of PURE. PURE uses a learning model that predicts a privacy label of a given image either *private* or *public*. The image is considered to be *private* if it belongs to the private sphere or contains objects that the user cannot share with the world and *public* otherwise.

PURE consists of two modules. The main module is the personalized learning module and serves as the core of the personal assistant. The purpose of this module is three-fold. First, using publicly annotated data, it learns to classify images as private or public. Second, it quantifies uncertainties in predictions, such that when it estimates a prediction to be highly uncertain, it can delegate the decision-making to the user. Three, it incorporates the user's expectations in privacy, as each user might have a different risk perception when it comes to how they would like to treat certain factors in the learning. The learning module 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 uses that to tune the personal assistant using the images of the user. This is important because privacy is inherently subjective and the publicly annotated dataset that is used for the learning module may not reflect the privacy expectations of the user. Moreover, due to the subjectivity, PURE might assign high uncertainty to some images. By fine-tuning using personal data, we aim to decrease the uncertainty that PURE might observe with some images. The second module is the decision-making module. When a user needs to make a privacy decision, this is the module that is invoked. This module obtains a prediction and an uncertainty value from the model. Each user defines for themselves when to let PURE make a decision and when they would want to be involved. By setting a threshold, a user can choose to decide on the privacy labels when the uncertainty is above the set threshold. Otherwise, the prediction of the model is assigned as the label.

Workings of PURE*:* OSN user has a personal assistant PURE. She can share her persona with her personal assistant. First, PURE has *Publicly Annotated Data* collected from different annotators. It learns privacy preferences using visual features in the *Learning Module* and produces *Model*. While learning, the user can share her persona that she can be sensitive about classifying private images as public or not. In this case, the personal assistant is risk-averse. Moreover, the user can share personal data that the user herself annotates allowing learning the user's privacy preferences. Then, PURE makes privacy decisions for its user's image (e.g. image) in the *Decision Module*. While making a prediction for each image, it also generates an uncertainty value for that prediction. To reach a privacy decision, PURE decides whether to use prediction results (i.e. *share* or *not share*) or to delegate the decision to the user (i.e. *delegate*) by comparing the uncertainty value with the threshold received from its user.

2.1.1. Learning Privacy Labels with Uncertainty

Evidential Deep Learning (EDL) [47] model quantifies predictions' uncertainty that is based on the Dempster-Shafer theory (DST) of evidence and Subjective Logic

(SL) [48, 49]. DST is the notion of a frame of discernment, which defines the set of all possible outcomes, along with a mathematical framework for reasoning under uncertainty and combining evidence from different sources. The main advantage of the DST is being able to represent ignorance and lack of evidence using belief functions and quantify the amount of uncertainty associated with different possible outcomes. SL builds upon the DST by expressing degrees of uncertainty through subjective opinions and introducing additional operators to handle uncertainty. When an opinion is formed over binary frames, it is known as a binomial opinion. When an opinion is on a frame that is larger than binary, it is known as a multinomial opinion. Each subjective opinion corresponds to a Dirichlet distribution, a conjugate prior to the categorical distribution. For a binary proposition (e.g. the image x is private), the subjective belief of a personal assistant for the truth of this proposition is represented as a binomial opinion, which corresponds to a Beta distribution —a special form of Dirichlet distribution [50, 51]. Since privacy classification is a binary classification task, a personal assistant's belief for an image to be private is represented as a binomial subjective opinion. A binomial subjective opinion for the classification can be represented as a Beta distribution. That is why, in this section, we will introduce the EDL using Beta distributions. Beta probability density function (pdf) is expressed as:

$$Beta\left(p|\alpha,\beta\right) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B\left(\alpha,\beta\right)}, \tag{2.1}$$

where B is the multivariate beta function [47] and $[\alpha, \beta]$ are the parameters of the Beta distribution. In Equation (2.1), p is the Bernoulli probability that the binary proposition is true, e.g. the probability that the image x is private. A personal assistant has a belief (b) for the proposition that *the image is private*, disbelief (d) for the same proposition, and an uncertainty (u) that represents the inability to classify the image accurately. The personal assistant's uncertainty about an image may be due to the noise in the image or the lack of training data with similar images. We can calculate these quantities as:

$$b = \frac{\alpha-1}{\alpha+\beta}, \quad d = \frac{\beta-1}{\alpha+\beta}, \quad \text{and} \quad u = \frac{2}{\alpha+\beta}, \tag{2.2}$$

where $b, d, u > 0$ and $b + d + u = 1$. Furthermore, $\alpha - 1$ and $\beta - 1$ are called the evidence for and against the proposition: *the image is private*. Let us note that u is maximized when $\alpha = \beta = 1$, corresponding to the uniform Beta distribution. We can also call them the evidence for the *private* and *public* categories in the classification of the image.

The Beta distribution provides a probability distribution over p —the probability that the given image is private. However, in classification tasks, we need a predictive categorical distribution to decide. For this purpose, we use the expected value of the Beta distribution, which is calculated as follows:

$$\bar{p} = \int_0^1 p\big(\frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}\big)dp = \frac{\alpha}{\alpha+\beta}. \tag{2.3}$$

The aforementioned calculations of belief masses and uncertainty are based on the parameters of the corresponding Beta distribution. In order to model belief masses and learn Beta distribution parameters, EDL modifies a vanilla neural network for classification by replacing its softmax layer with a non-negative activation function such as *ReLU*, *softplus*, and *exponential* functions. In our classification problem, we have two categories: *private* and *public*. Given a sample image x, we can use any neural network with two logits outputs: $o_0(x)$ and $o_1(x)$, one for each category. Then, we use the exponential function to calculate evidence for each category as follows: $e_{pub}(x) = exp(o_0(x))$ and $e_{pri}(x) = exp(o_1(x))$, which represent the evidence for the public and private categories, respectively. The Beta distribution parameters α and β for the classification of the image x are calculated as follows:

$$\alpha(x) = e_{pri}(x) + 1, \beta(x) = e_{pub}(x) + 1. \tag{2.4}$$

Let $y \in \{0, 1\}$ represent the category index of the sample image x. In standard neural networks for binary classification, the sigmoid function is used to calculate $p(x) = P(y = 1|x)$, i.e. the probability that x is from category $y = 1$. Then, the binary

cross-entropy loss is calculated as follows:

$$y \log \big(p(x)\big) + (1-y) \log \big(1-p(x)\big). \tag{2.5}$$

There are also other loss functions for classification, such as the Brier score, which is defined as

$$[p(x) - y]^2 + [1 - p(x) - (1-y)]^2. \tag{2.6}$$

The Brier score is a proper scoring function and is frequently used to measure the accuracy of probabilistic predictions. Unlike vanilla neural classifiers, we do not predict $p(x)$ directly, so we cannot directly use any of these loss functions. However, we predict its Beta distribution $Beta\big(p(x)|\alpha(x), \beta(x)\big)$; hence, we may calculate the expected loss by integrating out $p(x)$ in the classification loss of our choice. We can calculate the expected Brier score for privacy classification as follows:

$$\mathcal{L}(x,y) = \int_0^1 \big[p(x) - y\big]^2 + \big[1 - p(x) - (1-y)\big]^2 \frac{p(x)^{\alpha(x)-1}(1-p(x))^{\beta(x)-1}}{B\left(\alpha(x), \beta(x)\right)} dp(x), \tag{2.7}$$

which has the following closed-form solution:

$$\mathcal{L}(x,y) = [\bar{p}(x) - y]^2 + [1 - \bar{p}(x) - (1-y)]^2 + 2\frac{\bar{p}(x)(1-\bar{p}(x))}{\alpha(x) + \beta(x) + 1}, \tag{2.8}$$

where $\bar{p}(x)$ is the expectation of $p(x)$ and calculated as $\alpha(x)/\big(\alpha(x) + \beta(x)\big)$ using Equation (2.3).

$$\mathcal{L}(x,y) = \int_0^1 \Big[y \log \big(p(x)\big) + (1-y) \log \big(1-p(x)\big)\Big] \frac{p(x)^{\alpha(x)-1}(1-p(x))^{\beta(x)-1}}{B\left(\alpha(x), \beta(x)\right)} dp(x), \tag{2.9}$$

which has the following closed-form solution:

$$\mathcal{L}(x,y) = y\left(\psi\left(\alpha(x) + \beta(x)\right) - \psi\left(\alpha(x)\right)\right) + (1-y)\left(\psi\left(\alpha(x) + \beta(x)\right) - \psi\left(\beta(x)\right)\right), \tag{2.10}$$

where $\psi(.)$ is the *digamma* function. We also add a regularizing term $\mathcal{R}(x, y)$ to this loss. $\mathcal{R}(x, y)$ is defined as follows:

$$\mathcal{R}(x, y) = \lambda_t KL\left[Beta\big(p(x); \bar{\alpha}, \bar{\beta}\big) || Beta\big(p(x); 1, 1\big)\right], \tag{2.11}$$

where

- $t \geq 0$ is the index of the current training epoch,
- $\lambda_t = min\,(1.0, t/10)$ is the annealing coefficient,
- $KL[\cdot||\cdot]$ refers to the Kullback-Leibler (KL) divergence,
- $\bar{\alpha} = \alpha(x)^{1-y} = (e_{pri}(x) + 1)^{1-y}$,
- $\bar{\beta} = \beta(x)^{y} = (e_{pub}(x) + 1)^{y}$,
- $Beta\big(p(x); 1, 1\big)$ is the uniform Beta distribution.

Let us note that $\bar{\alpha}$ and $\bar{\beta}$ do not contain any evidence supporting the true category of the sample image. That is, $\alpha(x)^{1-y}$ becomes 1 if the image is private ($y = 1$) and $\beta(x)^{y}$ becomes 1 when the image is public ($y = 0$). As a result, the KL-divergence term is minimized when the network does not produce any evidence for the wrong category. Hence, this regularization term minimizes the evidence generated by the network for the wrong category and increases the predictive uncertainty for the misclassified samples [47].

Example 1: Let's assume that Alice is an OSN user. She has four different images. She needs to decide which images should be shared as public and which should be shared as private. Figure 2.2 represents an example for predicting privacy labels (such as public or private) of her images and quantifying uncertainty values for each prediction. In Figure 2.2, the first and the fourth images are public, and the other two are private. As shown in Figure 2.2, PURE predicts a label for each image as well as an uncertainty value. When producing an answer, it checks its uncertainty value and threshold to decide to answer with its current predicted label or delegate the decision to its user. If the threshold here is 0.7, it will put forward its predictions for images 2 and 3 (as



Figure 2.2. An example for predicting privacy labels of four images and quantifying uncertainty values for each prediction.

these have uncertainty values 0.1 and 0.5, respectively) and delegate image 1 and 4 to its user. With this setup, PURE would have correctly classified image 2 but image 3 would have been misclassified. It would have delegated image 1 that it would have failed on, but it would have also been delegated image 4 to the user that this image is correctly classified.

2.1.2. Personalizing Privacy

Since privacy is inherently subjective, it is important to incorporate the personal traits of the user into the decision-making. We consider three aspects of a user that should be factored into the decision-making: 1) perception of risk, 2) personal categorization, and 3) preference to be involved.

Perception of risk: While the personal assistant is making decisions, it is possible that it makes a prediction error. It is possible that for some users misclassifying a private image as public may lead to less desirable consequences than misclassifying a public image as private. For some others, there might not be a difference. Furthermore, the cost of different misclassifications may be significantly different for two different users. In order to avoid mistakes that are deemed risky for the user, the system needs to incorporate

the risk perception of the user into account.

Typically, vanilla neural networks do not differentiate this significant difference and consider all mistakes as equal. To overcome this, we introduce a user-dependent risk matrix, which is an asymmetric non-negative square matrix $R \in [0, \infty)^{2\times 2}$. Each value R_{ij} in R represents the user's cost when the classifier assigns an image from category i to the category j. There is no cost for the user for correct classification, hence $R_{ij} \geq R_{ii} = 0$.

There may be different ways of incorporating the user's risk of misclassification into the training of evidential classifiers. In this dissertation, we propose scaling misleading evidence in the KL-divergence term by modifying $\bar{\alpha}$ and $\bar{\beta}$ as follows:

$$\bar{\alpha} = \big(R_{01} e_{pri}(x) + 1\big)^{1-y}, \bar{\beta} = \big(R_{10} e_{pub}(x) + 1\big)^{y}. \tag{2.12}$$

This allows us to increase the KL divergence further when evidence for high-risk categories is produced. PURE gets R from its user and can learn how to generate evidence for each category based on the personalized cost of making misclassification. If the user is sensitive about classifying private images as public, the personal assistant also becomes sensitive and avoids generating evidence for the private category for equivocal and ambiguous images.

We can incorporate the user's risk of misclassification into the training of evidential classifier as follows:

(i) Scaling misleading evidence in the KL-divergence term by modifying $\bar{\alpha}$ and $\bar{\beta}$ as follows:

$$\bar{\alpha} = \big(R_{01} e_{pri}(x) + 1\big)^{1-y}, \bar{\beta} = \big(R_{10} e_{pub}(x) + 1\big)^{y}. \tag{2.13}$$

This allows us to increase the KL divergence further when evidence for high-risk

categories is produced.

(ii) Regularizing the amount of misleading evidence directly using the risk of misclassification.

$$(1-y)\bar{p}(x)R_{01}e_{pri} + y\big(1-\bar{p}(x)\big)R_{10}e_{pub}, \tag{2.14}$$

where $\bar{p}(x)$ is the predictive categorical distribution calculated as

$$\bar{p}(x) = \alpha(x)/(\alpha(x)+\beta(x)). \tag{2.15}$$

This term will be added to the aforementioned loss: $\mathcal{L}(x,y) + \mathcal{R}(x,y)$.



Figure 2.3. An example for predicting privacy labels of four images for a sensitive user and uncertainty values for each prediction.

Example 2: If for a user, there is no difference between the misclassification of private and public images, then PURE makes predictions as shown in Figure 2.2. On the other hand, assume that Alice is more sensitive about classifying a private image as public. By reflecting this in the R_{ij} score, PURE predicts privacy labels of images and quantifies uncertainties for each prediction as shown in Figure 2.3. Notice that the uncertainty values, as well as the predicted labels, have changed compared to Figure 2.2. With the uncertainty threshold still set to 0.7, PURE will delegate the same set of images (1 and 4) and answer images 2 and 3. Image 3 has been correctly classified this time. This

is a by-product of the fact that PURE chooses to classify more images as private to avoid the potential risk associated with classifying private images as public.

Personal Categorization: Another aspect of personalization is to understand what images a particular user finds private or public. One way of understanding this is to ask the user about privacy preferences. However, there is long-standing evidence that users are not good at articulating what they find private [13]. Moreover, their actions are not always in line with what they claim to be private [14, 15]. Thus, a better way of understanding what is private for a user is to utilize personal data: images that are labeled by the user herself.

PURE makes use of this to fine-tune the model it generates. After PURE is trained on publicly annotated data, the user's own labeled data is used to adjust the uncertainties in the model. An important contribution of this would be that the uncertainty in certain images drops such that model is more certain of its prediction.



Figure 2.4. An example for predicting privacy labels by fine-tuning using personal data and uncertainty values for each prediction.

Example 3: If PURE uses publicly available data and if Alice is a sensitive user about classifying a private image as public, PURE makes predictions in Figure 2.2 and 2.3, respectively. Moreover, if Alice shares her personal data that has been annotated by

her, PURE will predict privacy labels and uncertainty values for each prediction as shown in Figure 2.4. If uncertainty threshold is still 0.7, PURE will delegate image 1 to Alice and correctly classify image $2, 3$, and 4 (as these have uncertainty values $0.1, 0.2$, and 0.3, respectively). Since image 4 would have been correctly predicted with a lower uncertainty value, it would have not been delegated to the user. So, all the classifications would be correct this time, and PURE would ask its user less.

Preference to be involved: The final part of personalization is to understand how much a user wants to be involved in the decision-making. Recent work in HCI show that [24] while some users are happy to have privacy decisions taken by their privacy assistants on their behalf, some users would rather be in the loop. Moreover, this is not always a binary decision in the sense that with some decisions the user might want to be involved while with others she might not. We capture this preference to be involved in the decision making process explicitly using a threshold value θ. Whenever PURE is asked to label an image, in addition to a prediction, PURE also provides a level of uncertainty. When PURE has an uncertainty above θ, it delegates the decision making back to the user. Since θ can be configured by the user herself, it enables the user to select a level of involvement, where $\theta = 1$ would mean letting PURE do all the decisions, where $\theta = 0$ would mean overseeing all the decisions. During our experiments, we discuss having $\theta = 0.7$ as a working setting to capture user involvement only when PURE has high uncertainty.

2.2. Experimental Setup

We evaluate the performance of PURE in terms of its contribution to preserving privacy. Specifically, we aim to answer the first research question through its sub-research questions:

RQ1.1 Does PURE capture the ambiguity of privacy through its modeling of uncertainty, and by delegating ambiguous cases to the user, can PURE increase its privacy prediction accuracy?

RQ1.2 Does PURE adequately capture uncertainty and does it outperform existing models that capture uncertainty?

RQ1.3 Can PURE enable personalization of privacy by incorporating privacy risks and personal data of the user so that the accuracy is improved whilst fewer decisions are delegated to the user?

It is important to be able to answer RQ1.1 affirmatively because capturing the ambiguity of privacy is the key for PURE to choose when to consult its user. Ideally, uncertain images should be delegated to the user for a decision, and certain images should be answered by PURE. As a result, if we consider only the certain images PURE makes a prediction for, we would expect to obtain a higher accuracy than the overall accuracy. RQ1.2 investigates the dynamics between uncertainty and prediction errors, and questions whether alternative formulations of uncertainty such as an SNN or well-known uncertainty quantification methods such as MC dropout and Deep Ensemble would suffice. Finally, RQ1.3 explores if and to what extent personalization of PURE helps users, either in terms of the accuracy they obtain or the number of images they have to manually decide on.

Dataset: To evaluate our work, we selected a balanced subset of the PicAlert dataset [31]. PicAlert is a well-known benchmark dataset for the image privacy prediction problem and contains Flickr images that are labeled as *public* or *private* by external viewers. These images are the most recently uploaded images for a period of four months in 2010 and labeled by 81 users between 10 and 59 years of age with varied backgrounds. 17% of the images in this dataset have conflicting labels from annotators. We consider an image public if all the annotators have annotated it as public and private if at least one annotator has annotated it as private. Our subset contains 32000 samples labeled as *public* and *private*. It is split into *Train-set* and *Test-set* of 27000 and 5000 samples, respectively. Within the Train-set, we performed oversampling on the private images to align their count with that of the public images, as we had a total of 5667 unique private class images.

While previous research aims at increasing the accuracy of the privacy prediction, we additionally focus on how to quantify the uncertainty in these predictions and use it to improve the user's privacy in the face of automated decisions.

Metrics: We evaluate the performance of our approach using two main metrics: (i) success of the model in terms of standard metrics such as *Accuracy*, *F1*-score, *Precision*, and *Recall*; and (ii) ability of the model to quantify its predictive uncertainty, which allows for improvement in the success metrics in (i) if quantified correctly and accurately. We first evaluate our approach without considering personalization; hence the generated evidence is not weighted based on the perceived privacy risk of the user. Then, we extend our evaluations with the personalized risk matrices to see how our model adapts itself for users with different misclassification costs. Furthermore, we evaluate the model using personal data annotated by a user to observe how PURE adapts and then asks less from its user. To evaluate the quality of the uncertainty estimates, we calculate the accuracy of the model only on the test samples for which the model's uncertainty is less than a given uncertainty threshold between 0 and 1. When the uncertainty threshold is 1, all test samples are considered in computing the accuracy (and other metrics like precision and recall); however, when the threshold is reduced to 0.5, predictions with uncertainty values less than 0.5 are considered in the calculation of the success metrics.

Evaluation setting: We use models that are pre-trained on the ImageNet to extract features from images. We compare three popular deep architectures of CNNs: ResNet50, InceptionV3, and VGG16 in terms of their performance [52–54]. Table 2.1 shows the results of the comparison (*Accuracy*, *F1*, *Precision*, and *Recall*) of PURE using ResNet50, InceptionV3, and VGG16. ResNet50 and InceptionV3 pre-trained models yield better-performing models as compared to VGG16. Since residual connections enable the alleviation of the vanishing gradient problem, allowing for the successful training of networks, we choose *ResNet50* as our underlying architecture.

We use the *ResNet50* architecture as our base neural network and replace its last

	Accuracy	**F1**	**Precision**	**Recall**
ResNet50	0.89	0.89	0.89	0.89
InceptionV3	0.89	0.89	0.89	0.89
VGG16	0.73	0.72	0.8	0.73

Table 2.1. Performance of PURE using different pre-trained models ResNet50, InceptionV3, and VGG16.

layer (logits layer) with a densely connected layer with two outputs —one for each class (private and public). This architecture has 50 layers with residual connections. We implement our model using Tensorflow and initialize the network layers from the *ResNet50* model and train 15 epochs on the PicAlert dataset using Adam optimizer with a decaying learning rate initialized as $1e-5$. The *ResNet50* accepts images with dimensions ($224 \times 224 \times 3$), so we resize the images to these dimensions. PURE's implementation is available at `https://git.science.uu.nl/ayci0001/PURE`.

2.3. Results

We perform the following experiments with the mentioned dataset to answer our research questions.

Performance of PURE*:* We start with examining the Accuracy of PURE, where we configure PURE to provide a label no matter what the uncertainty is. We experiment with using the entire available training data as well as compare it to cases where the training data is smaller. Since a single OSN user shares a limited amount of data, it is crucial to perform well on small personal data, as also emphasized in [33, 55]. Furthermore, observing how PURE performs when trained with fewer data will enable us to understand to what extent PURE can yield satisfactory performance.

Table 2.2 shows the overall performance of PURE while training with different amount of data. For instance, when we use all data while training, PURE obtains an Accuracy of 0.89. PURE obtains an Accuracy of 0.87 for 25% of the Train-set. On

	Overall			
Usage %	**Accuracy**	**F1**	**Precision**	**Recall**
100	0.89	0.89	0.89	0.89
75	0.88	0.89	0.88	0.88
50	0.88	0.88	0.88	0.88
25	0.87	0.87	0.87	0.87
10	0.79	0.78	0.82	0.79
5	0.66	0.62	0.76	0.66
1	0.55	0.44	0.69	0.55

Table 2.2. Overall results for PURE as training samples are reduced.

the other hand, if PURE is trained on only 1% of data, the Accuracy decreases to 0.55. In addition to Accuracy, the performance of PURE monotonically decreases in terms of F1, Precision, and Recall while it uses less than 75% of the Train-set. Table 2.3 shows the performances of PURE for the private and public classes. For instance, PURE achieves F1 of 0.89 for each class when PURE uses all images in the training dataset. If a user has only 10% of the Train-set, PURE obtains F1 of 0.75 and 0.81 for the private and public class, respectively. In the extreme case of training with only 1% of the data, PURE exhibits poor performance in terms of F1, particularly for the private class. While the Precision for the private class samples remains consistent, a noticeable decrease in Recall is observed with a reduction in the amount of training data. PURE is good at correctly identifying the private samples among its predictions, even with reduced training data. However, the decrease in Recall indicates that PURE struggles to capture all the private class samples when it is not trained on sufficient private class data. This result is aligned with our expectation that because of the subjectivity of privacy, classifying private samples correctly can be challenging specifically while training on less data. Furthermore, due to the larger number of unique public class samples compared to the private class samples, PURE has an advantage that can learn subtle features specific to public images. These results are promising because it shows that even when there is limited training data, PURE can be useful.

	Private			Public		
Usage %	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**
100	0.89	0.91	0.87	0.89	0.87	0.92
75	0.88	0.91	0.87	0.89	0.86	0.92
50	0.88	0.92	0.84	0.89	0.85	0.92
25	0.86	0.91	0.81	0.87	0.82	0.92
10	0.75	0.92	0.63	0.81	0.72	0.95
5	0.49	0.94	0.34	0.74	0.6	0.98
1	0.19	0.85	0.11	0.68	0.52	0.98

Table 2.3. Results for the private and public classes of PURE at different training sample rates.

	Overall			
Delegation %	**Accuracy**	**F1**	**Precision**	**Recall**
0	0.89	0.89	0.89	0.90
10	0.92	0.92	0.92	0.92
25	0.95	0.95	0.95	0.95
50	0.97	0.97	0.97	0.97
75	0.99	0.99	0.99	0.99

Table 2.4. Overall results for PURE at prediction delegation rates 0%, 10%, 25%, 50%, and 75% based on uncertainty.

Recall that an important aspect of PURE is that it can calculate uncertainty. Next, we look at the relation between uncertainty and Accuracy to capture if PURE can represent uncertainty correctly. We have set up PURE so that it would delegate to its user when it is uncertain and thus is likely to make a mistake. Hence, ideally, when PURE delegates to its user, we would expect an improvement in the Accuracy of the remaining items.

Table 2.4 shows the overall performance of PURE with respect to different percentages of delegated predictions. When we do not delegate any predictions, PURE obtains an Accuracy of 89% (as was shown in Table 2.2). When we delegate only 25% of the most uncertain predictions, the results based on all performance metrics improve remarkably, e.g., the Accuracy, Recall, and Precision increase to 0.95. Similarly, when we delegate 75% of the most uncertain predictions, PURE achieves the highest performance of 0.99 in terms of all metrics. Thus, we observe that the delegated images are actually the ones that PURE would have made a mistake in.

	Private			**Public**		
Delegation %	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**
0	0.89	0.91	0.87	0.89	0.87	0.92
10	0.91	0.94	0.89	0.92	0.90	0.94
25	0.94	0.96	0.92	0.95	0.94	0.97
50	0.97	0.99	0.95	0.98	0.96	0.99
75	0.99	0.99	0.98	0.99	0.98	0.99

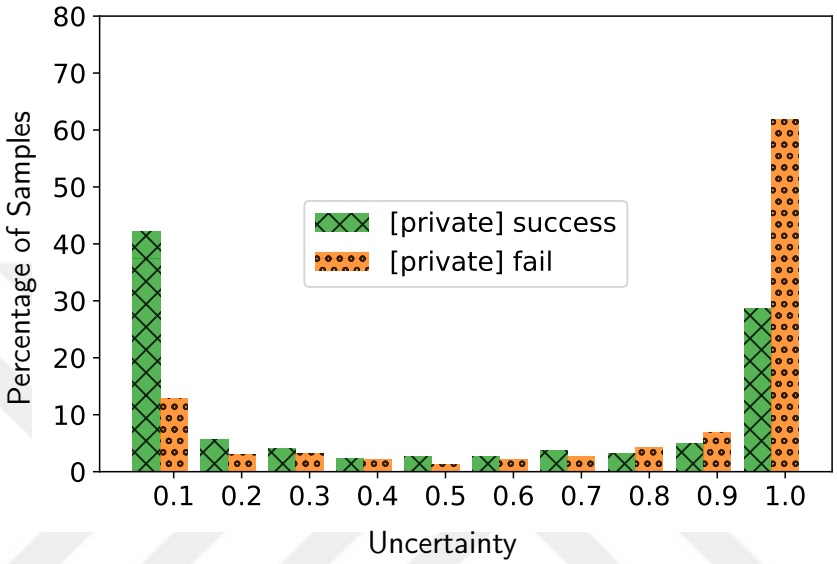
Table 2.5. Results for the private and public classes of PURE at various prediction delegation rates (0%, 10%, 25%, 50%, and 75%) based on uncertainty.

An important question is whether the same upward trend holds for both private and public classes. Table 2.5 shows the performance of PURE for each class. For instance, when PURE does not delegate any predictions, it obtains an F1 of 89% for both private and public classes. When PURE delegates only 25% of the most uncertain

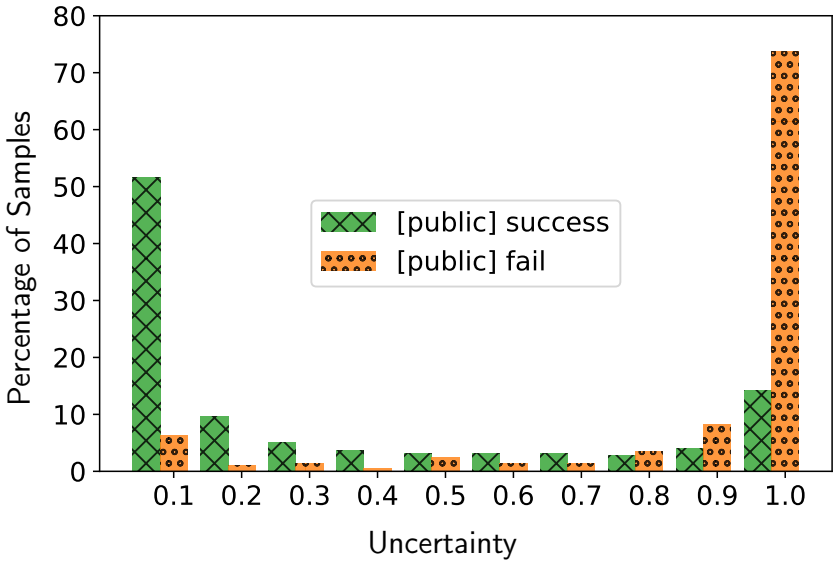
predictions, PURE improves F1s to 94% and 95% for the private and public classes, respectively. When it delegates 75% of the most uncertain predictions to its user, PURE yields the best performance with 0.99 F1s for both private and public classes, respectively. By increasing the number of delegated predictions, the performance of the model can be improved for each class significantly.

Another dimension to understand the link between uncertainty and making errors is to analyze what fraction of wrong predictions fall under different uncertainty rates. Figures 2.5(a) and 2.5(b) present the uncertainty histogram for the failed and successful privacy prediction of PURE, separately for the private and public classes. The failed and successful predictions in the uncertainty ranges of each class are shown as a percentage over each class itself. We have shown from these figures that correctly classified predictions for both the private and public classes tend to have low uncertainty values, indicated by their positioning on the left side. Conversely, misclassified predictions have higher uncertainty values, as positioned on the right side of each uncertainty distribution. For instance, 40% of the successful predictions in the private class have uncertainty values within the lowest uncertainty range (i.e. $[0, 0.1]$), while 63% of the failed predictions have uncertainty values within the highest uncertainty range (i.e. $[0.9, 1]$) in the uncertainty distribution. Similarly, for the public class, we see that 51% of the successful predictions have uncertainty values from the lowest uncertainty range, while 75% of the failed predictions have very high uncertainty values. So, we observe that failed predictions have higher uncertainty in general while successful predictions are more confident. This indicates that PURE is aware of its own ignorance and possible failures through its predictive uncertainty. When the most uncertain predictions are eliminated, its Accuracy improve drastically.

Similarly, Figure 2.6 plots uncertainty against Accuracy for PURE. The numbers on the particular points denote the ratio of test samples that are decided by PURE; the remaining samples are delegated back to the user because of the high uncertainty. The case when uncertainty is set to 1 is analogous forcing PURE to make all the privacy decisions, without delegating any case to the user. The Accuracy of PURE at this



(a) Private



(b) Public

Figure 2.5. Uncertainty distributions for the private and public classes.

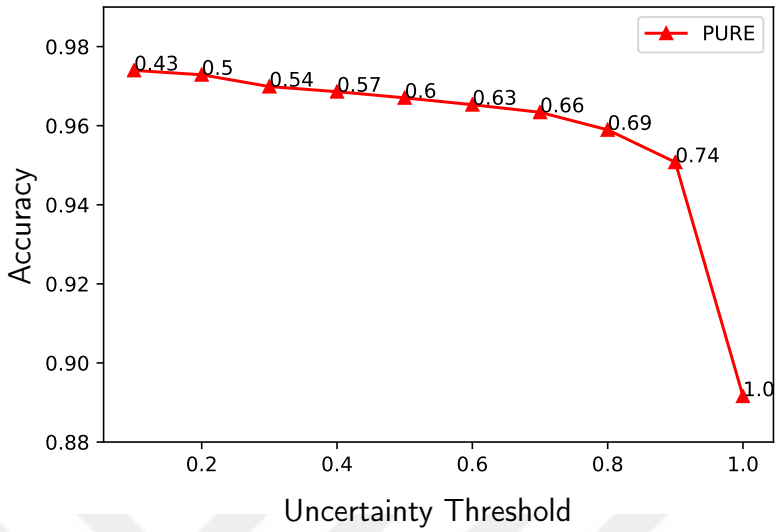


Figure 2.6. The change of Accuracy with respect to the uncertainty threshold.

stage is 89%. This is on par with the existing models in the literature that use the same dataset to predict privacy labels [32]. The more interesting cases are the ones where the uncertainty is high so that the PURE decides that there is too much uncertainty to answer and delegates them to the user. For example, for uncertainty threshold 0.4, 57% of the test samples can be decided by PURE, leading to an Accuracy around 0.97. For uncertainty threshold 0.8, 69% of the test samples can be decided with PURE, leading to an Accuracy around 0.95. This shows, as RQ1.1 asks, that PURE can capture the privacy ambiguity and can delegate such cases to its user.

Comparison with Alternative Networks: PURE calculates the uncertainty of its predictions and exploits it to refrain from making wrong privacy decisions for its users. In order to understand the effect of PURE in its calculations of uncertainty, we compare it to alternative predictive uncertainty models; Monte Carlo (MC) dropout [56] and Deep Ensemble [57], which depend on the class probabilities predicted by the neural network as well as a regular Standard Neural Network (SNN). We implement an SNN, MC dropout, and Deep Ensemble with two softmax outputs using the same *ResNet50* architecture with PURE. In order to measure the uncertainty of standard deep classifiers, the entropy of their predictions have been used after normalizing it to have an

uncertainty value between 0 and 1 [47, 58].

To have a meaningful comparison for uncertainty quantification, we use the normalized entropy as a proxy for the uncertainty for the PURE and SNN, MC dropout, and Deep Ensemble models. For PURE, we use the expected probabilities defined in Equation (2.3) to calculate the entropy. The entropy for the class probabilities p and $(1-p)$ is calculated as $-[p\log p + (1-p)\log(1-p)]$; then normalized by dividing to $\log 2$, which is the maximum entropy for the binary classification.

Gal and Ghahramani propose MC dropout method that represents model uncertainty using dropout in neural networks at test time [56]. We add dropout layers after each non-linearities and set the dropout rate as 0.05[1]. We train a model, obtain the desired number of different predictions and take the average of five predictions for each class. Lakshminarayanan *et al.* propose ensemble-based method, called Deep Ensemble that quantifies predictive uncertainty [57]. We use the Brier score, Equation (2.6), as a proper scoring rule as the training criterion, train five models with the same architecture, and take the average of predictions.

Figure 2.7 plots the test Accuracy of the PURE, SNN, MC dropout, and Deep Ensemble models for different normalized entropy thresholds (for delegating predictions). PURE outperforms the SNN, MC dropout, and Deep Ensemble models at almost every point. Its Accuracy is higher than that of alternative models even when there are no delegated predictions and the disparity significantly increases as the predictions are filtered based on their entropy. When we select entropy threshold as 0.4, SNN achieves 95.6% Accuracy using 51% of the data. For the same threshold, 53% of the data used by MC dropout, and the Accuracy value is 95.2%. Deep Ensemble obtains an Accuracy of 93% using 88% of the data, whereas PURE achieves 97.8% of Accuracy for 48% of the predictions at the same threshold. Our observation is consistent with the literature, where the deep neural networks are criticized as being overconfident, hence misleading, when they make mistakes [58]. One important aspect to note here is

[1] 0.05 yields the best performance among $\{0.01, 0.1, 0.25, 0.5\}$.

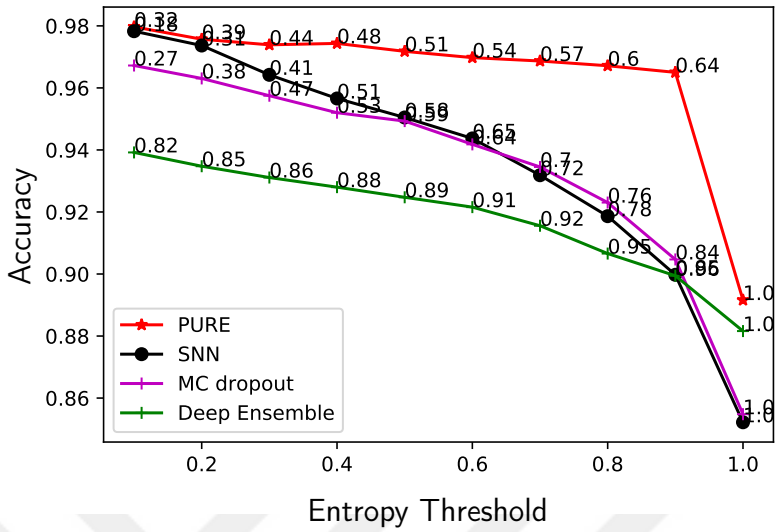


Figure 2.7. The change of Accuracy for different models with respect to different entropy thresholds.

the distribution of the data over various entropy thresholds. In principle, we want to use the entropy threshold to decide if a decision will be delegated to the user. Consider PURE in Figure 2.7. When entropy is 0.1, PURE will only classify 32% of the data and delegate the remaining to the user. While this is a large percentage to delegate, it comes with the advantage of 98% Accuracy. If PURE sets its entropy to 0.8, then it will classify 60% of the data and still yielding an Accuracy of 98%. When it chooses to classify all the data, then the Accuracy will drop to 89%. Contrast this ability to configure based on entropy to Deep Ensemble. With Deep Ensemble, even when the entropy is set to 0.1, the personal assistant will classify 82% of the data itself, with low flexibility in delegating the choices to the user. Next, we study the Accuracy changes of these models based on their data usage.

Figure 2.8 plots the F1 for varying delegated data fraction percentages. This shows how the F1 changes for the *public* and *private* classes when PURE, SNN, MC dropout, and Deep Ensemble models delegate certain percentages of their most uncertain predictions based on the entropy. The F1 scores of PURE and Deep Ensemble models are better than SNN and MC dropout for each class and the gap is bigger for the pri-

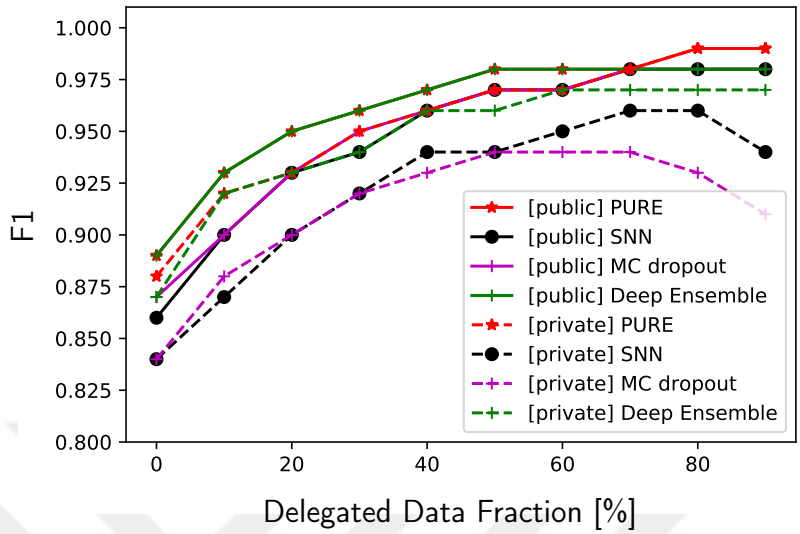


Figure 2.8. F1 scores for the private and public classes relative to the percentage of delegated decisions.

vate class. PURE outperforms Deep Ensemble for both classes when the rate is higher than 0.7. The F1 of PURE improves further and reaches 0.99 for both private and public classes. However, the F1 of the private class for SNN and MC dropout decreases when the most uncertain 80% of the data is neglected and the remaining most certain 20% is used for the calculation of the F1. The decrease in the F1 for the private class in Figure 2.8 indicates that SNN and MC dropout is overconfident while PURE can exploit its well-measured uncertainty to avoid wrong privacy decisions. With randomization tests [59], we can show that the improvements of PURE over existing models are statistically significant (p-value < 0.05). We answer RQ1.2 positively such that PURE outperforms SNN, MC dropout, and Deep Ensemble by expressing uncertainty better.

Personalized Misclassification Risk: As each user is different when it comes to privacy, they may each also have a significantly different cost for the misclassification of private content. In this section, we demonstrate the flexibility of PURE to adapt to users' perceived risk of such mistakes and its ability to avoid them by refraining from making privacy decisions when uncertain.

Persona	**Non-Sensitive**	**Semi-Sensitive**			**Sensitive**
Risk Values	$R_{01} = 1$ $R_{10} = 1$	$R_{01} = 1$ $R_{10} = 3$	$R_{01} = 1$ $R_{10} = 5$	$R_{01} = 1$ $R_{10} = 7$	$R_{01} = 1$ $R_{10} = 10$
[Overall] Accuracy	0.89	0.89	0.90	0.90	0.90
[Overall] Recall	0.89	0.89	0.90	0.90	0.90
[Private] Recall	0.86	0.87	0.89	0.90	0.91
[Public] Recall	0.92	0.91	0.90	0.89	0.89

Table 2.6. Results for different risk personas.

In order to capture users' varying degrees of sensitivity to the misclassification of private content, we construct so-called privacy personas. Each persona has its own distinct risk matrix R which reflects how sensitive they are to misclassifications. Overall, we consider three categories of personas: *non-sensitive*, *semi-sensitive*, and *sensitive*, with *semi-sensitive* subsequently comprised of three sub-categories with different sensitivity levels. One potential way to determine a user's privacy persona could, for example, be through administering a short questionnaire.

A non-sensitive user has the same perception of risk for the misclassification of private and public images, i.e. $R_{01} = R_{10} = 1$. This means that for a non-sensitive user, the regularization (KL) term in the loss of PURE weighs the evidence for private and public classes any differently. On the other hand, for semi-sensitive users, misclassifying a private image as public is more unacceptable, i.e. $R_{01} = 1$ and $R_{10} = \{3, 5, 7\}$. The latter reflects varying degrees of sensitivity *within* the semi-sensitive category. Finally, we also have sensitive users for whom misclassifying a private image as public is ten times more unacceptable compared to non-sensitive user, i.e. $R_{01} = 1$ and $R_{10} = 10$. This means PURE is penalized significantly more for making a wrong prediction in the private class as opposed to the public class.

Table 2.6 shows the overall Accuracy, as well as Recall results for the overall, private class, and public class, across various personas. For the non-sensitive persona, the results are as before. For the sensitive persona, PURE prefers to classify content as

private over public, when in doubt. This behavior increases the number of predictions for the private class for the sensitive user (i.e. private Recall increase from 0.86 to 0.91). While doing so, it does not sacrifice its overall Recall and Accuracy, with both increasing from 0.89 to 0.90. As we have shown from the table, the Recall for the private class improves significantly at the cost of having a lower Recall for the public class. Our results indicate that by increasing R_{10}, we increase private Recall, thus classifying more images as private, and as a result, we obtain a lower Recall for public images. Notice that the R value belongs to a user and can be adjusted as needed.

Data Personalization: PURE delegates a decision to its user when uncertain about its prediction. Ideally, we would like to minimize the number of times this happens while keeping the Accuracy high. The personalization is meant to serve this purpose. In order to see if this is indeed achieved, we need to perform a comparative analysis where in the first round, no personal data are used and in the second round, the personal data are added.

- *Round I:* Train a model without using personal data.
- *Round II - Personal:* Tune a trained model in *Round I* using personal data annotated by a user.
- *Round II - Random:* Tune a trained model in *Round I* using data annotated by others.

Figure 2.9 plots the delegation rate of PURE for the aforementioned training rounds for three different users. These are the top three users in terms of number of annotated images, i.e. the users who have annotated the most images out of all annotators. Hence, they have the most data available for tuning. The data shows that simply training the model without the use of any personal data (Round I) results in relatively high delegation rates for all three users. Introducing a second round (Round II - Random) where the model is tuned with more (random) data already yields an improvement in terms of the delegation rate for the users two and three. More data

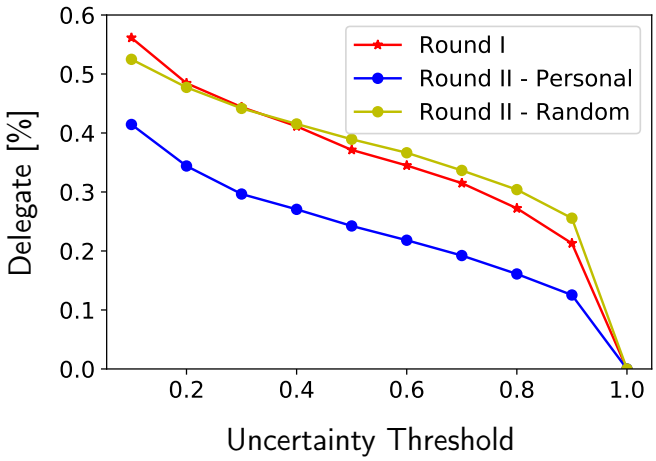
	Round I	**Round II - Personal**	**Round II - Random**
User 1	0.32	0.19	0.34
User 2	0.30	0.21	0.27
User 3	0.34	0.22	0.29

Table 2.7. Ratios of predictions whose uncertainty values are greater than 0.7 (θ).

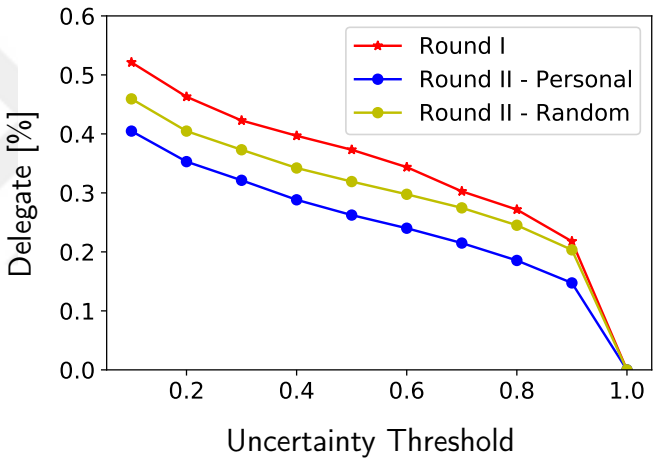
appears to result in a better tuned model even when it does not contain data from the user in question. However, when PURE is tuned with personal data (Round II - Personal), performance can be improved further (i.e. even less images have to be delegated to users). Indeed, we observe that PURE tuned with personal data performs best for all three users as it results in the least amount of images delegated to the user. It appears that personalization can result in a better understanding of the user's privacy preferences, likely making it easier for PURE to determine whether an image is considered public or private for this specific user. In essence, the use of the personalization module can allow PURE to be more certain about its predictions, and consequently delegate less images to the user.

Table 2.7 shows the percentage of delegated images per user for each round, where the uncertainty threshold is 0.7. For instance, for the first user, PURE delegates 32% and 34% of test samples after *Round I* and *Round II - Random*, respectively. However, only 19% of images have been delegated to the user when we tune the trained model using personal data in Round II. In light of these results, we answer RQ1.3 positively: PURE can adjust its behavior based on the personal risks and expectations of its user as well as help reduce the number of decisions delegated to the user.

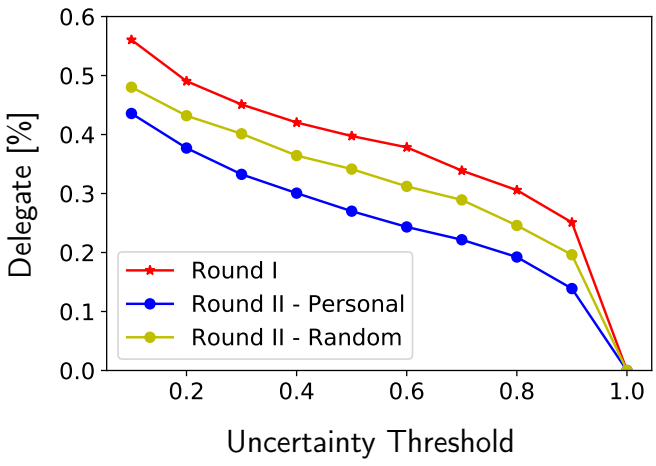
An interesting aspect to note is that the number of images used to personalize can affect the behavior of the data personalization approach. Given the limited number of annotators with a predefined set of images, we currently cannot study such questions but it would be useful to provide bounds to guide users to personalize the assistant even further.



(a) User 1



(b) User 2



(c) User 3

Figure 2.9. The delegation rate by round for different uncertainty thresholds.

2.4. Conclusion

We have proposed how PURE helps its user to make privacy decisions by recommending privacy labels (private or public) for a given image. PURE is uncertainty-aware as it captures the ambiguity of privacy using uncertainty modeling, and delegates decisions for ambiguous cases to its user. Our experimental results show high accuracy for PURE, even when the user is not consulted at all. We observed that most of the images identified as uncertain are those it would have missclassified if not delegated to the user. As a result, we answer RQ1.1 affirmatively since PURE is able to increase its privacy prediction accuracy through delegating uncertain images to the user. Furthermore, our comparison with alternative networks in the literature shows that PURE captures uncertainty well and outperforms alternative models such as SNN, MC dropout and Deep Ensemble. Therefore, RQ1.2 is answered positively. Finally, PURE allows for personalization of privacy as it is capable of making a privacy decision by incorporating the user's risk of misclassification and using personally labeled data. Through its personalization, PURE is also unobtrusive as it consults its user only when it is uncertain. Hence, we also answer RQ1.3 positively in light of our results that show PURE can indeed adjust its behavior based on the personal risk and expectations of its user, and is consequently able to decrease the number of delegated decisions by fine-tuning using personal data. In conclusion, the development of PURE allows us to positively answer this dissertation's Research Question 1. This chapter has been published in [60, 61].

3. IMAGE PRIVACY PREDICTION USING TEXTUAL FEATURES

As users share content, it is necessary for them to think about for whom the content is meant and how to configure the privacy settings [30]. An important category of content is images. Recent work helps users categorize whether a given image is private or not [32]. This can be useful to help users avoid the unintended sharing of private images on social networking sites. However, privacy is *personal* and *subjective*, and therefore an assistant ought to provide personalized answers as to whether an image is private or not [33, 60]. In Chapter 2, we have introduced and evaluated PURE, which helps its OSN users in making privacy decisions by considering the ambiguity of privacy predictions, users' privacy personas, and personal data. To ensure the adoption of these privacy assistants by end-users, it is crucial to establish trust. One important path to induce such trust is through explanations [34]. By utilizing semantic information such as descriptive keywords and topics, users can gain an understanding of the features used in the decision-making process. These keywords or tags can be provided by the user or automatically generated by a tool. Leveraging these generated tags allows us to extract latent topics that provide additional meta-information about the content. Topics capture thematic aspects associated with the images, thereby enhancing the classifier's ability to make contextually-aware privacy predictions.

This chapter investigates how we can make use of textual features (i.e. tags and topics) for image privacy predictions. All images are represented by tags (i.e. descriptive keywords). Figure 3.1(a) displays an example image and Figure 3.1(b) shows 20 different tags generated by Clarifai API[2]. A qualitative assessment of the example image shows it is captured well by the generated tags.

First, to gain a deeper understanding of the underlying factors, we employ a term weighting method to extract informative features. In cases where a tag is ab-

[2]https://clarifai.com/clarifai/main/models/general-image-recognition



01	child	11	portrait
02	girl	12	toy
03	education	13	classroom
04	people	14	room
05	school	15	class
06	son	16	group
07	indoors	17	elementary school
08	boy	18	cute
09	teacher	19	fun
10	family	20	adult

(a) Image (b) Generated Tags

Figure 3.1. Example image and its generated tags by Clarifai.

sent during the training phase, we replace the unseen tag with the most semantically similar tag that exists. We assess the impact of tags generated by different CNN models. Additionally, we analyze the importance of word-embedding models in enhancing the accuracy of privacy predictions. As we delve into the analysis of these tags, valuable insights emerge, specifically grouping them based on their semantic similarity. Then, we utilize a topic modelling technique to extract latent topics from tags associated with images. Each image can be associated with a single topic or multiple topics. Our experiments demonstrate that the topic-based approach achieves remarkable performance in learning privacy preferences of users. By leveraging the insights obtained from tags and the extraction of latent topics, we enhance the classifier's ability to make contextually-aware privacy predictions and get ready to build an explainable system.

Organization. Section 3.1 explains how to identify and generate tags for making privacy predictions. Section 3.2 investigates how to extract topics from an image's generated tags. Section 3.3 evaluates the representativeness of the topics and shows experimental results of the prediction performance. Section 3.4 presents our conclusions.

3.1. Understanding Underlying Factors

We employ the Term Frequency-Inverse Document Frequency (TF–RF) term weighting method [62] to extract information from a collection of textual data. This

method calculates the weight of a *tag* j in an *image* i as follows:

$$tfrf_{i,j} = tf_{i,j} \times log\left(2 + \frac{r_{i,j}}{s_{i,j}}\right), \tag{3.1}$$

where

- $tf_{i,j}$ is the number of occurrences of *tag* j in *image* i,
- the constant value 2 is the base of this logarithm function,
- $r_{i,j}$ denotes the number of occurrences of *tag* j in all public images,
- $s_{i,j}$ denotes the number of occurrences of *tag* j in all private images

In this dissertation, we refer to the method as *TF–RF* and denote an input for a classifier as *tfrf*. Every image is associated with 20 distinct tags. Therefore, when utilizing all tags of the images, the term frequency value ($tf_{i,j}$) becomes $\frac{1}{20}$ for each tag. To prevent potential division by zero errors, such as when a tag is absent in all private images, we introduce the addition of epsilon (ϵ) as follows:

$$tfrf_{i,j} = tf_{i,j} \times log\left(2 + \frac{r_{i,j} + \epsilon}{s_{i,j} + \epsilon}\right), \tag{3.2}$$

where $\epsilon \in (0, 1)$.

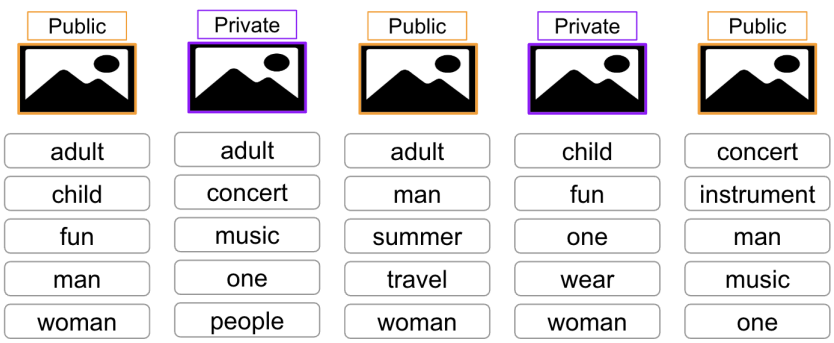


Figure 3.2. An example of images and their tags.

Example 1: Assuming Alice has five images, as depicted in Figure 3.2, the public (orange) ones include the first, third, and fifth images, while the private (purple) ones

consist of the second and fourth images. Each image is associated with five distinct tags. Examples of the r and s matrices can be seen in Figure 3.3 and 3.4, respectively. Additionally, Figure 3.5 illustrates an example of a relevancy matrix constructed using Equation (3.2) for Alice's images. In these matrices, each row corresponds to an image, and each column corresponds to a unique tag. Initially, we construct the r and s matrices by counting the number of occurrences of tags in public and private images, respectively. Subsequently, we generate an input matrix for the classifier, which captures the relevance of public and private images based on the constructed r and s matrices using Equation (3.2). The indices in the r and s matrices have values greater than or equal to zero, indicating the tag's occurrence count. On the other hand, the elements in the relevancy *tfrf* matrix are real numbers. It is worth noting that certain tags may be common across multiple images. For instance, the tag "adult" appears in the first, second, and third images, which are public, private, and public, respectively. Thus, it occurs twice in the public images and once in the private image. When a tag does not pertain to an image or belongs to a private image, corresponding value in the r matrix is set to zero. There are two scenarios where the values in the r and s matrices become zero. First, if a selected tag does not belong to an image, it will have a zero value. Second, if a tag from an image is missing in either the public or private images, it will result in a zero value. We obviously understand that when a tag of an image has a non-zero value, it provides more information for predicting the privacy labels of the image.

	adult	child	concert	fun	instrument	man	music	one	people	summer	travel	wear	woman
Public	2	1	0	1	0	3	0	0	0	0	0	0	2
Private	2	0	1	0	0	0	1	1	0	0	0	0	0
Public	2	0	0	0	0	3	0	0	0	1	1	0	2
Private	0	1	0	1	0	0	0	1	0	0	0	0	2
Public	0	0	1	0	1	3	1	1	0	0	0	0	0

Figure 3.3. r matrix for Example 3.1.

In order to construct a *tfrf* relevancy matrix, the system works as follows: the algorithm of the TF–RF approach initializes r, s, and *tfrf* matrices. It gets data which

	adult	child	concert	fun	instrument	man	music	one	people	summer	travel	wear	woman
Public 	1	1	0	1	0	0	0	0	0	0	0	0	1
Private 	1	0	1	0	0	0	1	2	1	0	0	0	0
Public 	1	0	0	0	0	0	0	0	0	0	0	0	1
Private 	0	1	0	1	0	0	0	2	0	0	0	1	1
Public 	0	0	1	0	0	0	1	2	0	0	0	0	0

Figure 3.4. s matrix for Example 3.1.

	adult	child	concert	fun	instrument	man	music	one	people	summer	travel	wear	woman
Public 	0.28	0.22	0.22	0.22	0.22	2.52	0.22	0.22	0.22	0.22	0.22	0.22	0.28
Private 	0.28	0.22	0.22	0.22	0.22	0.22	0.22	0.18	0.14	0.22	0.22	0.22	0.22
Public 	0.28	0.22	0.22	0.22	0.22	2.52	0.22	0.22	0.22	2.30	2.30	0.22	0.28
Private 	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.18	0.22	0.22	0.22	0.14	0.28
Public 	0.22	0.22	0.22	0.22	2.30	2.52	0.22	0.18	0.22	0.22	0.22	0.22	0.22

Figure 3.5. *tfrf* matrix for Example 3.1.

contains images and their tags for both Train and Test sets. r and s matrices store the number of occurrences of every unique tag in all public and private images as described in Equation (3.2). The expected output is a *tfrf* relevancy matrix which is later given to the classifier.

3.1.1. Tag Identification

The TF–RF method computes the frequency of tag occurrences in both public and private images. A classifier utilizes the TF–RF method to predict labels for given images. However, if a tag in the Test set is not present in the Train-set, the classifier lacks information regarding the privacy status associated with that tag. Instead of disregarding the unseen tag, we leverage the vector representation of tags.

Each tag can be represented as a vector by utilizing pre-trained word embedding models, as demonstrated below:

- BERT[3]
  - BERT [63] is a pre-trained contextualized word embedding model which is published by Google. We use *bert-embedding* library in Python.
- Word2Vec[4]
  - Word2Vec [64] is a popular word embedding models which uses shallow a neural network. We use pre-trained *Google News corpus* which has (3 billion running words) word vector model (3 million 300−dimensional English word vectors).
- GloVe[5]
  - GloVe [65] is a popular unsupervised learning algorithm for word representation which is developed at Stanford University. *Common Crawl* pre-trained model which has 42B tokens, 1.9M words, and all words are uncased as tags of images and each of them has 300−dimensional vectors.

In this approach, the algorithm suggests tags by comparing them to similar tags in the Train-set using a cosine similarity metric. When encountering an unseen tag, we substitute it with the most similar tag in the Train-set. The cosine similarity between two attribute vectors is illustrated as follows:

$$cosine_similarity(A,\ B) = \frac{A \cdot B}{\|A\|\,\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}. \tag{3.3}$$

Example 2: Let's consider the scenario where Alice intends to share an image with everyone, which includes 20 different tags. One of these tags is *soccer*. However, *soccer* is not present in the public images, resulting in a value of zero for this tag. In such cases, we recommend replacing the unseen tag with the most similar tag from the public tag list using vector representations of tags and the *cosine-similarity* metric. For instance, let's assume that *football* is a tag present in the public tag list of the Train-set. The cosine similarity value between the vectors representing *soccer* and *football* is 0.81. Among all the tags in the public tag list, *football* exhibits the highest similarity

[3]https://pypi.org/project/bert-embedding/
[4]https://code.google.com/archive/p/word2vec/
[5]https://nlp.stanford.edu/projects/glove/

values to *soccer*. Consequently, we select *football* tag and replace the original *soccer* tag with *football*.

3.1.2. Tag Generation

To assess the importance of i) tag generation ii) tag representation models, we employ a Random Forest classifier [66,67]. We evaluate the performance of the classifier in terms of Accuracy, F1, Precision, and Recall. For our evaluation, we select a subset of the PicAlert dataset, consisting of 7000 images that are randomly divided into 3500 public and 3500 private images. It is worth noting that all the images are either 100% public or private.

Importance of Tag Generation Models: Before evaluating the prediction performance of the classifier (Random Forest algorithm) based on feature representation, it is essential to observe the impact of the tags used. For this purpose, we examine how the performance of the classifier changes when using different tag generation methods.

	Accuracy	**F1**	**Precision**	**Recall**
VGG	0.65	0.61	0.69	0.69
ResNet	0.68	0.66	0.71	0.73
GoogleNet	0.69	0.66	0.71	0.74
AlexNet	0.69	0.66	0.71	0.75
Clarifai	0.86	0.86	0.90	0.83

Table 3.1. Comparison of privacy prediction performance of the classifier by using different tag generation methods such as VGG, ResNet, GoogleNet, AlexNet, and Clarifai.

We obtain performance results for using tags generated by *CNN models* [52, 54, 68, 69] and *Clarifai API*. Specifically, we utilize tags generated by CNN models as described in [70, 71]. We remove 64 images from the dataset due to the absence of

tags for images in CNN models. We utilize the Clarifai API[6] in order to automatically generate a set of 20 different tags for each image. The architecture of the general image recognition model of Clarifai is CNNs such as InceptionV2. Their model is trained on over 20 million images and uses $10,000$ concepts to identify objects in images and videos. As shown in Table 3.1, Clarifai generates more informative tags for the classifier, resulting in an accuracy of 0.86. This outcome is aligned with our expectation because Clarifai generates tags that are more readable and meaningful for humans compared to tags generated by deep learning models, such as "pjs", which might not make sense to human interpreters. Therefore, we utilize Clarifai API to generate image tags for our study.

Importance of Word-Embeddings: Besides the TF–RF approach, alternative input preparation methods for the classifier (Random Forest algorithm) involve employing word embedding models, namely BERT, GloVe, and Word2Vec. Specifically, for GloVe and Word2Vec models, we construct a vector for each image by combining all tag vectors by taking the *minimum* value for each dimension[7].

Model	**Accuracy**	**F1-score**	**Precision**	**Recall**
BERT	0.82	0.80	0.84	0.81
GloVe	0.84	0.83	0.85	0.83
Word2Vec	0.84	0.83	0.85	0.84
TF–RF	0.86	0.86	0.90	0.83

Table 3.2. Comparing privacy prediction performance obtained through various input preparation methods, including the TF–RF, as well as word embedding techniques like BERT, GloVe, and Word2Vec.

Table 3.2 shows the prediction performances of the classifier for various word embedding (i.e. BERT, GloVe, and Word2Vec), and the TF–RF methods. We observe that the classifier achieves satisfactory prediction results using the word embedding models (i.e. accuracy values are $0.82, 0.84$, and 0.84 for the BERT, Glove, and Word2Vec

[6]https://clarifai.com/clarifai/main/models/general-image-recognition
[7]minimum yields the best performance among {summation, maximum, and mean}.

models, respectively). On the other hand, the classifier outperforms BERT, GloVe, and Word2Vec when we construct feature vectors using the TF–RF method and replace tags that are not present in the Test set with the values of the semantically closest tags (i.e. accuracy values are $0.82, 0.84, 0.84$, and 0.86, respectively). When using the TF–RF method without making recommendations for missing tags, we obtain an accuracy of 0.84.

Through our observations, we have discovered that leveraging textual features, such as tags generated by Clarifai API, significantly contributes to the learning of privacy labels. Furthermore, we have explored the use of semantic similarity as a valuable resource in cases where tags are missing. By grouping tags into topics based on their semantic similarity, we can obtain high-level contextualized semantic information. By doing so, we can enhance the representation of the images and potentially build up explanations using less textual input.

3.2. Extracting Topics

Topic modelling is a technique that discovers latent topics within a collection of textual information. It allows us to extract distinct topics from an image's generated tags. We employ the widely used Non-negative Matrix Factorization (NMF) topic modelling technique [72].

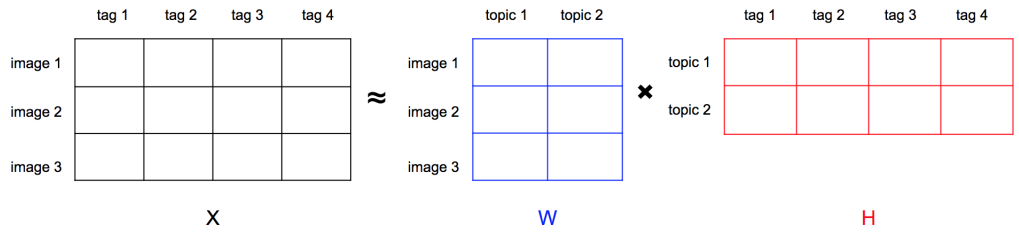


Figure 3.6. Non-negative Matrix Factorization concept.

NMF is an approximation to factorize a non-negative image-tag matrix X into non-negative matrices W and H as illustrated in Figure 3.6. The W (features) matrix denotes the degree to which an image belongs to a topic and the H (components) ma-

trix denotes the degree to which a tag belongs to each topic. The W and H matrices are initialized randomly. The NMF algorithm runs iteratively until it finds W and H matrices that minimize the Frobenius norm of the matrix, that is, $\|X - W \times H\|_F$. NMF is suitable for interpretability (components are non-negative) and works better and faster for short texts (a set of tags) compared to alternatives such as Latent Dirichlet Allocation (LDA) [73]. In this study, we make use of the term weighting method and we specifically employ the Term Frequency - Inverse Document Frequency (TF-IDF) model to measure the presence of tags. TF-IDF allows us to transform tags into numerical vectors in order to construct an *image-tag (X)* matrix. We learn the tag vocabulary from the images and then transform the images into a *image-tag (X)* TF-IDF matrix. Each row of the matrix corresponds to an image, and each column represents a unique tag from the images. The value in each cell of the matrix represents the TF-IDF weight of the tag in the corresponding image. We build the NMF model for a given number of topic (k) values, which generates an *image-topic (W)* matrix and a *topic-tag (H)* matrix. Note that the number of topics k is an important parameter to set, which we explain next.

3.3. Evaluation

Dataset: To evaluate the representativeness of the topics extracted using NMF with the images, we trained a Random Forest classifier where the images are represented as TF-IDF vectors of these topics. The subset of the PicAlert dataset we have chosen for our study comprises $32K$ images, consisting of $27K$ and $5K$ images for the *Train-set* and *Test-set*, respectively.

Metrics: We evaluate the performance (accuracy, f1-score, precision, and recall) on the Test set separately for private and public image classes. This is critical as the consequences of misclassifying a private image could be more severe than misclassifying a public image.

Evaluation of Topics: Topics should be meaningful and interpretable for humans. One

way of realizing this is to ensure that the topics are *coherent*, which means topics should be relatively different from each other (i.e. distinct) whilst images within a certain topic should be described by similar keywords. Hence, we can measure coherence based on two different criteria as follows:

(i) Intra-topic similarity: The average semantic similarity between all pairs of the most associated N tags in the same topic. Hence, it is a measure of how similar a topic's tags are to each other. For instance, the most associated three tags with *Topic Nature* and *Topic Child* are $\{tree, park, wood\}$ and $\{child, baby, fun\}$, respectively. In our example, both topics have high intra-topic similarity.

(ii) Inter-topic similarity: The average semantic similarity of the most associated N tags from different topics. Hence, it is a measure of how much overlap there is between different topics' tags. For example, the associated tags with *Topic Nature* and *Topic Child* are clearly not overlapping, meaning they have a low inter-topic similarity.

We can evaluate the quality of the explored topics by calculating these two measures of similarity. For good topic modelling, we want to 1) maximize the intra-topic similarity, thus ensuring the topic is well-defined and the associated tags are closely related to each other, and 2) minimize inter-topic similarity, thereby ensuring the topics are distinct and the tags within each topic are not closely related to the tags in the other topics.

In the NMF model, we set the number of topics based on the model's performance in terms of *coherence*. While calculating intra-topic and inter-topic similarity, each tag is represented by word embedding vectors, namely word2vec [64]. We represent tags as $300-$dimensional vectors of the word2vec model trained on Google News when calculating coherence[8]. The similarity between two tag vectors is measured by the *cosine similarity* metric. Semantically similar tags tend to be close to each other in the semantic space. Intra-topic similarity values for 10 and 20 topics are 0.18 and 0.20,

[8]https://code.google.com/archive/p/word2vec/



(a) Nature (b) Child (c) Performance (d) Business (e) Fashion

Figure 3.7. Tag clouds for Topics Nature, Child, Performance, Business, and Fashion.

respectively. Note that the cosine-similarity values between two vectors for this model are generally low. For instance, the similarity between "person" and "people" is 0.51 and "tree" and "park" is 0.23. Hence, $k = 20$ is better in terms of intra-topic similarity as the tags within a given topic are more related to each other. Additionally, inter-topic similarity values for 10 and 20 topics are 0.48 and 0.43, respectively. In this case, $k = 20$ is again a better fit as the lower value indicates more distinct, i.e. better segregated, topics. Thus, we set the number of topics to 20.

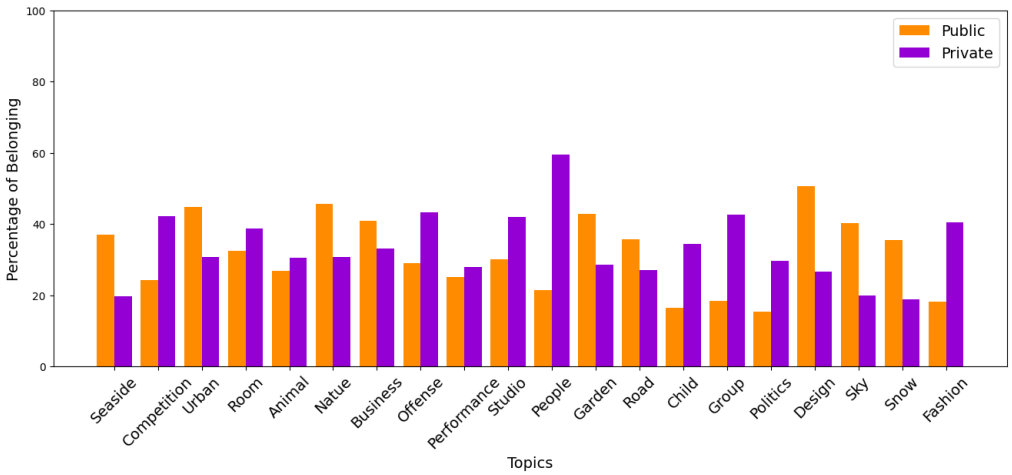


Figure 3.8. Percentage of occurrence of each topic in private and public images.

The topics extracted from NMF are nameless and thus not quickly understandable. Hence, we name the 20 topics manually since the names will be used in the generated explanations. Figure 3.7 shows tag clouds of the top 20 tags for five different topics. The font size indicates relative significance, i.e. the most descriptive tag is displayed as the largest. For instance, the top five descriptive tags of the topic *Nature* are $\{tree, park, wood, nature, outdoors\}$. Figure 3.8 shows the share of each topic associated with private and public images. Some topics such as *People* are associated more frequently with the private class whereas others like *Sky* are more often associated with

the public class. However, despite some topics being more frequently associated with a specific class, the topics ultimately do not have an explicit class to which they belong.

Results: Our findings demonstrate that the classifier achieves a precision of 87% and 89%, recall of 90% and 87%, and f1-score of 89% and 88% for private and public classes, respectively. Overall, the classifier yields an accuracy of 88% on the Test set, indicating the NMF-extracted topics are effective for privacy prediction. It achieves state-of-the-art performance for image privacy prediction compared to existing approaches [32, 60]. In light of the results, we observed that we accurately predict privacy labels by leveraging textual features.

3.4. Conclusion

In this chapter, we have investigated how we utilize textual features from semantic information while making privacy predictions. Text-based approaches can allow us to understand features that have been influential on the decision. First, we adopt the TF–RF weighting method to extract tags. The TF–RF captures the local importance of the tag within tags of a certain image and its global relevance within the tags collection of all images. When a tag is not present in the Train-set, we recommend semantically similar tags for this tag. Through our experiments, we have shown the importance of tag generation and using word-embedding methods. We have also observed that when a tag in the Test-set is present in both the public and private classes of the Train-set, and its occurrence is similar in frequency across both classes, there is a potential for the image to be misclassified. Moreover, the recommendation of semantically similar tags provided us valuable insights into grouping tags to establish meta-level topics. By grouping tags as topics and subsequently reducing the feature space, it becomes possible to build explanation systems. Motivated by this observation, we extract latent topics which provide insights into contexts present in the tags collection of all images. We envision that each image can be associated with single or multiple topics, each with varying degrees of relevance. By leveraging these topics, we achieve state-of-the-art privacy prediction performance. This chapter has been published in [74, 75].

4. EXPLAINABLE PRIVACY ASSISTANT

Privacy assistants can collaborate with users, offering valuable support in the management of their privacy-related responsibilities [23, 24]. When users engage in content sharing, thoughtful consideration regarding the intended recipients and the required configuration of privacy settings becomes essential [30]. Images are an important category amongst the various types of content shared online. Recent research has focused on enabling users to classify whether an image should be regarded as private or public [32]. This classification can prove beneficial in preventing the unintentional sharing of private images on OSNs. Given the personal and subjective nature of privacy, where users' perceptions of privacy may vary, it becomes imperative for an assistant to deliver personalized responses concerning the privacy status of images [33, 60]. To ensure the widespread acceptance of privacy assistants among end users, establishing trust becomes crucial. One effective approach to cultivate such trust is through the provision of explanations [34]. Consequently, we tackle a novel challenge related to the interplay between explanations and privacy: how can a privacy assistant explain the rationale behind classifying a particular piece of content as either private or public?

Explainable Artificial Intelligence (XAI) suggests approaches that aid in the comprehension of why and how a machine learning (ML) model arrives at its prediction. There are various explanation methods with respect to visual and textual explanations [36–39]. For the image privacy prediction task, an example of a visual explanation can be highlighting the most important region in the image for the target class, whereas a textual explanation can be a generated text such as "if a guitar had not been in the image, the image would not have been public" [39]. A visual explanation can quickly point out a potential privacy concern, however, it can be overwhelming for users to understand underlying mechanisms since it may present many intricate details [36, 37, 40–42]. Szymanski *et al.* [43] suggest that the combination of visual and textual explanations improves understanding for non-expert (without any software or domain knowledge) users. We propose an explanation method that provides a visual

representation accompanied by a textual description.

In Chapter 3, we have investigated the use of textual features (i.e. tags and topics) for making privacy predictions. While these approaches perform well in making privacy predictions, the indispensability of trust in personal assistants becomes evident as an adaptation of end-users. This chapter presents a privacy assistant PEAK that decides whether a given image is private or not, and explains this decision to its user as well as other privacy assistants. To the best of our knowledge, this is the first privacy assistant that can explain why an image is considered private or public. PEAK utilizes automatically generated tags of images, explores latent topics from the tag sets, classifies images based on image-topic associations, and ultimately generates human-understandable explanations. An explanation involves presenting one or more topics related to an image in a visual format, highlighting significant tags associating the image to a certain topic, and providing a textual description detailing the connections between the topics. PEAK derives these explanations from a well-known image dataset for privacy where images are labeled as public or private [31]. We perform a user study to gauge if participants actually find the generated explanations useful and what factors of the explanation or the image affect users' understanding of the model decision. Furthermore, privacy assistants may delegate some privacy decisions to their users to avoid failures when they are uncertain. This generally improves the accuracy of privacy decisions [60] but leads to an increase in cognitive load for the user. In the cognitive load theory, Sweller *et al.* [76] state that individuals have limited cognitive resources available in working memory which is responsible for information processing. A higher number of statements stored in working memory will subsequently result in an increase in cognitive load. By combining PEAK with PURE, we are able to reduce the number of images delegated to users, without compromising the accuracy of the privacy predictions, thus helping to reduce the cognitive load for users.

Organization. Section 4.1 explains the mechanism of our privacy assistant in detail. Section 4.2 presents how to identify explanation categories and generate explanations. Section 4.3 evaluates our explanation model with a user study and shows experimental

results of PEAK working in combination with privacy assistants. Finally, Section 4.4 concludes our work.

4.1. Preserving Privacy with PEAK

PEAK is a privacy assistant that explains privacy labels by generating human-understandable explanations. The generated explanations revolve around topics (collections of related keywords) as these are intuitive and easier to understand for users. An image can belong to multiple topics with varying degrees of relatedness, and these relationships can help understand why an image is considered to be public or private. For end users, the most important aspects of an explanation are simplicity and relevance. Hence, the explanation is visual in nature and accompanied by a short explanatory text describing the connections between the topics and the image. The explanation is focused on the most important features of the image and associated topics, rather than being an exhaustive analysis. In order to achieve this, only a subset of the most relevant topics associated with an image are used to explain why the image is classified as public or private. Ultimately, PEAK helps OSN users by allowing them to understand why images have been identified as private or public. Moreover, PEAK can aid privacy assistants in enhancing their performance on making privacy decisions (Section 4.3). Figure 4.1 shows the workflow of PEAK.

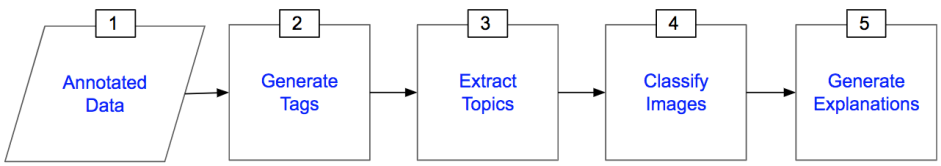


Figure 4.1. Workflow of PEAK.

PEAK is composed of the following five stages:

(i) The starting point for PEAK is a set of labeled public and private images. The data could come from a user's own history of personal online images (i.e. both shared and not shared), or from a publicly available dataset of labeled images.

(ii) The second stage involves assigning tags, i.e. descriptive keywords, to each image. The generated tags are in plain language such as "tree" or "baby", and they can be provided by users themselves or generated automatically by a tool such as *Clarifai*.

(iii) The next step is using the set of labeled and tagged images to perform topic modelling, which is a technique used to extract latent (i.e. hidden) topics from textual information (i.e. the tags). Each image is associated with one or more topics and topics are constructed using two criteria. First, tags within a single topic should be semantically related, meaning images associated with the same topic should be described by similar tags. Second, topics should be semantically distinct from each other. In other words, there should not be too much overlap between different topics. Finally, in order to make it easier to understand for the user, we named the generated topics manually. For example, the topic with the tags "tree", "parks", and "outdoors" was named Nature.

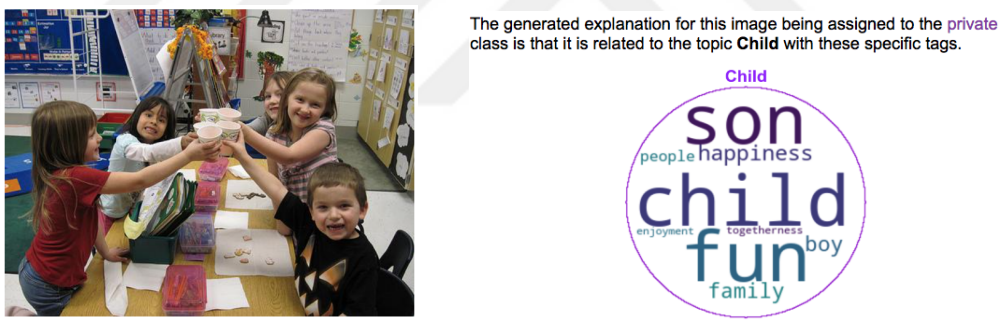
(iv) Once topics have been generated, the next stage involves training a tree-based ML algorithm for binary image classification. In this case, the generated topics serve as features of the images and the algorithm uses image-topic relationships to predict whether a new image should be classified as public or private. Additionally, the contributions of each topic to the privacy decision (i.e. positive or negative effect) are computed.

(v) The final step is identifying explanation categories for images, which essentially are image profiles with certain characteristics in terms of topic contributions. For instance, there may be a single dominant topic that pushes the prediction overwhelmingly in one direction. Based on the topics' contributions to privacy prediction, images are assigned to one of four explanation categories we identified: Dominant, Opposing, Collaborative, and Weak. Finally, for each explanation category, there is a distinct textual and visual explanation pattern based on the image's topics and the relationships between topics.

For end-users, it is important that any explanation generated by PEAK is simple and easy to understand, i.e. no technical knowledge should be required. Ultimately,

our aim is not to explain the technical process of the classifier to the end-user. Additionally, since many users do not actually read long, complicated privacy policies, the explanation should be visually understandable and supported by a short text.

Based on these constraints, we propose an explanation as to why an image is considered public or private using a set of *topics* belonging to the image. Each image can have one or more topics and these topics are shown as a circle which is labeled with the topic name. Additionally, we identify one or more *tags* linking this image to each topic and denote them in the corresponding topic circle. The visual representation aims to explain that the image is public or private because it is described by the displayed topics and tags. Furthermore, the visualization is augmented with a short explanatory description using a predetermined language structure. The text is thus supplementary and does not provide additional information.



(a) Image (b) Generated Explanation

Figure 4.2. An example image annotated as private and its generated explanation by PEAK.

Figure 4.2(a) displays an example of a private image. Figure 4.2(b) presents the explanation generated by the PEAK method. The explanation shows the image is classified as private because it is associated with topic *Child* and its related tags. Whilst in this example only one topic is displayed, there can be multiple topics contributing to the prediction. In that case, PEAK visually represents only the most relevant topics and explains the relation between these different topics.

4.2. Generating Explanations from Topics

We start walking through the steps outlined in Figure 4.1. We use a balanced subset of the PicAlert dataset that comprises 32000 images, consisting of 27000 and 5000 images in the Train-set and Test-set, respectively. We have explained how to assign tags to images and extract topics from the image-tag relation (Step $1-3$) in Chapter 3.2. We use the Random Forest algorithm to make image privacy predictions (Step 4). The question may arise whether certain topics might be more often associated with public or private images. If so, that may make it relatively easy to explain the privacy predictions. As we have already shown in the Figure 3.8, some topics being more frequently associated with a specific class, so the topics ultimately do not have an explicit class to which they belong. Therefore, a topic by 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 a topic's class.

4.2.1. Computing the Contribution of Topics using the TreeExplainer Model

The TreeExplainer [38] is the implementation of the SHAP (SHapley Additive exPlanations) approach [37], which can be used to understand how an ML model arrived at its prediction. The TreeExplainer model provides the computation of local explanations based on Shapley values in polynomial time. The model computes the contribution of each feature to a prediction, taking into account the interactions between features using tree-based models such as the Random Forest algorithm. In this study, each feature corresponds to a topic. Not all topics have an equal contribution to a class prediction: a topic can push the prediction higher (positive SHAP value) or lower (negative SHAP value), and their magnitude can differ. An ML model concludes its prediction by taking into account the contribution of each topic. This is useful in interpreting how the model works. One way to create explanations would be to display all these values to the user. However, as the number of topics increases, showing them all to the end user would be cumbersome and confusing. Therefore, we start with the TreeExplainer output and then interpret it to make it understandable for the end user.

We are interested in identifying topics that are useful in explaining the content of the image at hand. For example, for a given image, a large positive SHAP value might be assigned to a topic because the image is related to that topic. But, it might also be the case that a large negative value is assigned to a topic that is unrelated to the image. The latter shows the model made a decision based on the fact that the image did not exhibit the properties associated with this topic. While useful for the model designers, this information is difficult to understand for the end-users. Hence, we need to carefully decide how to use the SHAP values when creating the explanations.

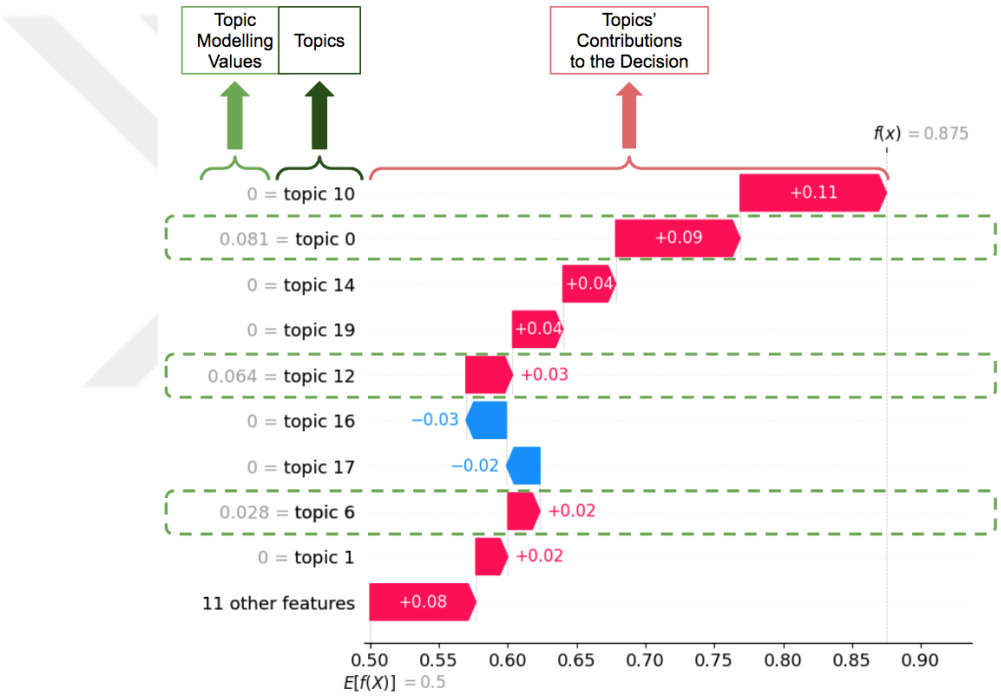


Figure 4.3. Example output of the TreeExplainer model.

Figure 4.3 is an example output of the TreeExplainer model for a given image [38]. The TreeExplainer model plots topics and their contributions on the prediction. Since, we have binary (i.e. public or private) classification problem, the expected model output value is 0.5 as shown on the x-axis and adding all contributions to this value gives the prediction probability of being in the certain class which is 0.875 in this example. We modify the output of the TreeExplainer model employing the topic modelling result. That is, we filter topics based on their association with the image. On the left column of the figure, we have shown only "topic 0", "topic 12", and "topic 6" are associated with this image. They are listed in the order of their contribution to the

decision, with values of 0.09, 0.03, and 0.02 (on the right column) respectively. Hence, the TreeExplainer model shows the contribution of every topic used in the prediction. However, some topics are not associated with the image and are thus not used for the explanation. Therefore, we exclude these topics when generating the explanation, thereby focusing only on the most relevant topics.

4.2.2. Identifying Explanation Categories of Images

Our privacy assistant generates human-understandable explanations by first reducing the number of topics in the TreeExplainer model output. We bring structure to the explanations through the use of categories, thereby reducing complexity and making it easier to understand for users. We make a distinction between explanations generated for end users and those for agents as they have different needs. For end users, our explanation consists of two components, namely a short, explanatory text and a visual representation of the most important topics associated with the image. Utilizing both a textual and a visual representation ensures the explanation is intuitive and easier to understand. The image's category remains a key input though as each category has its own distinct explanatory text and visual representation. In contrast, for agents, the sole explanation is the image category as no other information is necessary.

First of all, following extensive manual work and analysis of the topics' contributions to the decisions, three distinct image categories were identified, namely *Dominant*, *Opposing*, and *Collaborative*. Images not definitively belonging to any of the three categories are grouped in a separate category called *Weak*. Second, our approach generates explanatory text for each explanation based on the explanation category. That is, the explanatory text function has different patterns for different explanation categories in order to explain the visual representation and the relations between topics. Third, our explanation approach provides a visual representation to indicate whether the image is public or private based on the topics and tags displayed. Depending on the image's explanation category, it visualizes one or more topics with the related image tags in a circle labeled with the name of the topic(s).

The four explanation categories are further explained below and example explanations are given.

Dominant: An image belongs to the *Dominant* category when the contribution of one topic is decisive for the class prediction. In this case, a topic makes a relatively high contribution compared to all other topics of the image. To determine whether a contribution is considered decisive for the prediction, a lower bound (db) of 0.7 is used. Hence, when there is a topic whose contribution exceeds this threshold, the image is considered to belong to the Dominant category.

If the image is categorized as *Dominant*, the generated explanation for a user contains only the dominant topic in both the text pattern and the visual representation. The pattern of the explanatory text for the Dominant category is as follows: "The generated explanation for the image being assigned to the public/private class is that it is related to the *topic x* with the specific tags". Figure 4.2 shows an example of a *Dominant* image identified as private by annotators. For this example, the contribution of the topic Child is 8.6 times larger than the topic with the next greatest contribution.

Opposing: The topics associated with an image do not always conclusively point to whether the image should be classified as public or private, i.e. opposing signals exist. Hence, an image belongs to the *Opposing* category if it has topics whose contributions to a class prediction are in opposite directions and whose magnitudes are comparable *and* sufficiently large. The latter is the case if the opposing contributions both exceed the lower bound (ob) of 0.2, i.e. the contributions need to be sufficiently large. When an image has opposing topics whose contributions are in opposite directions, making a class prediction can be difficult. Hence, when this is the case, the explanation to the user should clearly indicate this and the pattern of the explanatory text for the Opposing category is therefore as follows: "Even though it is related to the *topic x* with the specific tags below (which signals the public/private class), it is also related to the *topic y* and for that reason, it is classified as private/public". Figure 4.4 shows an example image identified as public by annotators. The generated explana-

tion explicitly mentions the opposing topics, i.e. Child and Design. The reason why the image has ultimately been classified as public is that Design's contribution higher outweighs Child's contribution pushing the decision lower.



Figure 4.4. Example image annotated as public and its generated explanation with the topics Design and Child (Opposing category).

Collaborative: An image belongs to the *Collaborative* category if *enough*, though not *all*, topics are *decisively* pushing the decision in the same direction. Unlike the Dominant category, there is no single decisive topic. Instead, there are multiple topics whose combined collaborative contributions are decisive in classifying the image. If the sum of the collaborative contributions exceeds the Collaborative lower bound *cb*, equal to 0.8, the image belongs to the Collaborative category. Finally, the presence of a topic (or topics) pushing the decision in the opposite direction (of the collaborative direction) does not necessarily mean the image does not belong in the Collaborative category. As long as those opposing contributions are relatively minor, the image is still considered to be Collaborative rather than Opposing. The explanatory text of

the Collaborative category includes the top three collaborative topics as follows: "The generated explanation for the image being assigned to the public/private class is that it is related to the *topics x, y,* and *z* with these specific tags". Figure 4.5 shows an example image identified as private. The generated explanation provides the topics People, Fashion, and Room (with relevant tags), which all push the decision to be private.

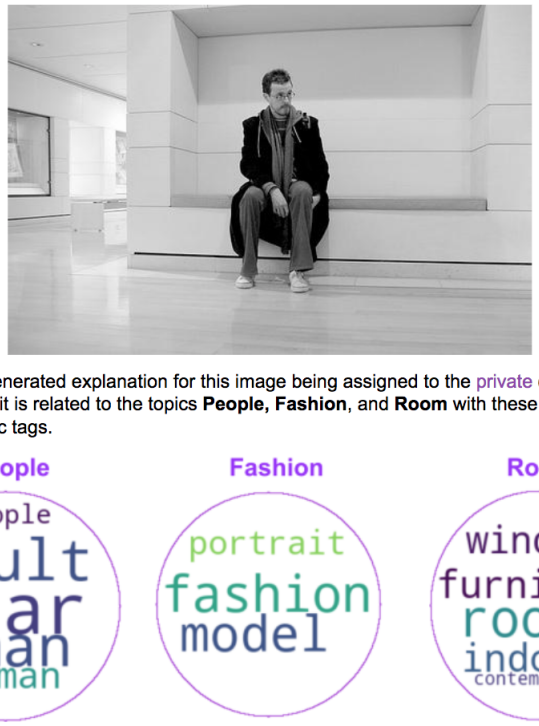


Figure 4.5. Example image annotated as private and its generated explanation with the topics People, Fashion, and Room (Collaborative category).

Weak: Finally, it is also possible an image belongs to many topics with minor contributions. Hence, it would not fall into any of the above three categories and its class can therefore not be explained as clearly as the others. We call this category *Weak* and generate an explanation containing only the top contributing topics, i.e. topics with relatively small contributions are not shown. In doing so, our aim is to generate explanations with the most relevant and influential topics for the decision. The explanatory text pattern for the Weak category, shown in Figure 4.6, is the same as the explanatory text for the Opposing category. Since none of the topics in this figure have a large contribution to the decision, no decisive decision can be made.

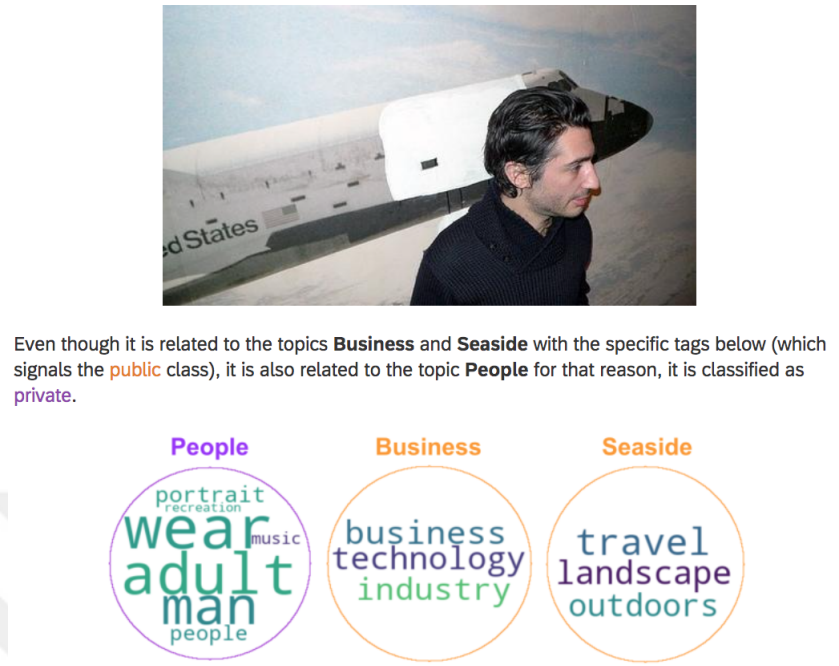


Figure 4.6. Example image annotated as private and its generated explanation with the topics People, Business, and Seaside (Weak category).

4.2.3. Computing Explanations

Our method identifies four different explanation categories by utilizing the contributions of topics on the decision.

Algorithm 1 shows how our proposed approach categorizes an image, gives appropriate topics and tags to be used in the user explanation, and ultimately generates an explanation (specific to each category) why the image is classified as public or private. In this algorithm, T is the number of topics associated with an image. *topic_vector* is a vector containing the SHAP values of the associated topics for the image. Additionally, we normalize these values by dividing each value by the sum of the absolute values of all image topics and storing the normalized values of *topic_vector* in *norm_vector*. In a final step, the algorithm creates a sorted vector (*sorted_vector*) containing the first N topics of the *norm_vector*, sorted in descending order.

db, ob, and *cb* are the respective lower bounds, as previously mentioned in the

Algorithm 1: Generate Explanation for an Image

Input: img, a given image,

$T \in \mathbb{N}$, the number of topics,

$topic_vector \in \mathbb{R}^T$, stores the SHAP values of the associated topics for img,

$norm_vector \in \mathbb{R}^T$, stores the normalized values of the topic_vector,

$sorted_vector \in \mathbb{R}^N$, stores the indexes of the first N topics sorted by $norm_vector$ in descending order,

$\{db, ob, cb\} \in [0, 1]$, lower bounds of Dominant, Opposing, and Collaborative categories, respectively,

TP and TG, topics and tags spaces, respectively,

i2t : $\mathbb{Z}^+ \to TG$, index to topic function,

et: $\sum^* \times 2^{TP} \times 2^{TP} \to \sum^*$, explanatory text function,

im2t : $I \to 2^{TG}$, image to tag function,

tp2t : $TP \to 2^{TG}$, topic to tag function.

Output: explanation: $\sum^* \times \sum^* \times 2^{TP \times TG}$

explanation $\leftarrow (\emptyset, \emptyset, \emptyset)$

$c_sum^+, c_sum^- \leftarrow 0, 0$

$c_topics^+, c_topics^- \leftarrow [], []$

$o_topics^+, o_topics^- \leftarrow [], []$

if $norm_vector\,[sorted_vector\,[1]] \geq db$ **then**

  $d_topic \leftarrow i2t(sorted_vector\,[1])$

  $d_tags \leftarrow im2t(img) \cap tp2t(d_topic)$ $d_topic_tags \leftarrow (d_topic, d_tags)$

  explanation $\leftarrow$

  $(\text{“}dominant\text{”}, et(\text{“}dominant\text{”}, \{d_topic\}, \emptyset), d_topic_tags)$

else

  for $n = 1$ ***to*** N **do**

    if $topic_vector[sorted_vector\,[n]] > 0$ **then**

      $c_sum^+ \leftarrow c_sum^+ + norm_vector[sorted_vector\,[n]]$

      $c_topics^+ \leftarrow c_topics^+ \cup \{i2t(sorted_vector\,[n])\}$

      if $norm_vector[sorted_vector\,[n]] \geq ob$ **then**

        $o_topics^+ \leftarrow o_topics^+ \cup \{i2t(sorted_vector\,[n])\}$

14 **else if** $topic_vector[sorted_vector\,[n]] < 0$ **then**

15 $c_sum^- \leftarrow c_sum^- + norm_vector[sorted_vector\,[n]]$

16 $c_topics^- \leftarrow c_topics^- \cup \{i2t(sorted_vector\,[n])\}$

17 **if** $norm_vector[sorted_vector\,[n]] \geq ob$ **then**

18 $o_topics^- \leftarrow o_topics^- \cup \{i2t(sorted_vector\,[n])\}$

19 **if** $|o_topics^+| \in \mathbb{Z}^+$ *and* $|o_topics^-| \in \mathbb{Z}^+$ **then**

20 $o_topics_tags^+ \leftarrow$
$\bigcup_{i=1}^{|o_topics^+|} \{(o_topics^+[i], im2t(img) \cap tp2t(cf_topic^+[i]))\}$

21 $o_topics_tags^- \leftarrow$
$\bigcup_{i=1}^{|o_topics^-|} \{(o_topics^-[i], im2t(img) \cap tp2t(cf_topic^-[i]))\}$

22 $explanation \leftarrow (\text{“}opposing\text{”}, et(\text{“}opposing\text{”}, o_topics^+, o_topics^-),$
$o_topics_tags^+ \cup o_topics_tags^-)$

23 **else if** $c_sum^+ \geq cb$ **then**

24 $c_topics_tags \leftarrow$
$\bigcup_{i=1}^{3} \{(c_topics^+[i], im2t(img) \cap tp2t(cl_topic^+[i]))\}$

25 $explanation \leftarrow$
$(\text{“}collaborative\text{”}, et(\text{“}collaborative\text{”}, c_topics^+, \emptyset), c_topics_tags)$

26 **else if** $c_sum^- \geq cb$ **then**

27 $c_topics_tags \leftarrow$
$\bigcup_{i=1}^{3} \{(c_topics^-[i], im2t(img) \cap tp2t(c_topic^-[i]))\}$

28 $explanation \leftarrow$
$(\text{“}collaborative\text{”}, et(\text{“}collaborative\text{”}, \emptyset, c_topics^-), c_topics_tags)$

29 **else**

30 $w_topics \leftarrow i2t(sorted_vector\,[:3])$

31 $w_topics_tags \leftarrow$
$\bigcup_{i=1}^{3} \{(w_topics[i], im2t(img) \cap tp2t(w_topics[i]))\}$

32 $explanation \leftarrow (\text{“}weak\text{”}, et(\text{“}weak\text{”}, w_topics, \emptyset), w_topics_tags)$

category definitions, with regard to deciding whether a given image belongs to the Dominant, Opposing, or Collaborative category. A number of functions are needed in order to generate category-specific output explanations, with TP and TG as respectively the topics and tags spaces, i.e. all explored N topics and associated tags. The function $i2t$ returns topic names from the indexes of $sorted_vector$, e.g. the first element is topic Child. The function et is the explanatory text function that creates the text pattern used in the explanation. It uses as input the image's category and its explanation topic(s), which are the most relevant topics for the prediction. For instance, only the Dominant topic for Dominant images. The $im2t$ function returns the tags associated with a given image whilst the $tp2t$ function returns the tags associated with a given topic. The functions together are needed to ultimately find the common tags associated with *both* the image and a certain topic. Finally, Algorithm 1 features as output *explanation*, which consists of three components, namely the image's category name, the generated explanatory text from the et function and the explanation topic(s) with associated tags.

Several variables are initialized (line 1), starting with *explanation*. Additionally, in order to be able to determine whether an image belongs to the Collaborative category, the algorithm initializes two variables, namely the total sum of the contributions of topics pushing the prediction higher (i.e. c_sum^{+}) or lower (i.e. c_sum^{-}). Secondly, two variables used to store the explanation topics for Collaborative images are initialized, respectively c_topics^{+} and c_topics^{-} for higher- and lower pushing topics. Finally, the algorithm initializes two variables containing lists of topics that push the prediction higher (i.e. o_topics^{+}) or lower (i.e. o_topics^{-}). These variables are used to assess whether a given image belongs to the Opposing category.

Algorithm 1 first checks whether the image is in the Dominant category. When there is a single topic whose contribution is disproportionately high compared to the other topics, the image is deemed to be Dominant. Hence, if the normalized value of the first topic in the $sorted_vector$ (i.e. the topic with the largest contribution) is equal to or greater than the Dominant's bound, the image is considered Dominant (line

2). Then the $i2t$ function subsequently returns the Dominant topic name (d_topic, line 3) whilst its associated tags d_tags are returned from the intersection of the $im2t$ and $t2t$ functions (line 4). The two are then stored in d_topic_tags (line 5). Finally, the algorithm outputs the explanation using the generated inputs (line 6).

If the image is not Dominant (line 7), variables for the Collaborative and Opposing categories are constructed by cycling through each topic up to N (line 8) and based on the direction of the topic's contribution to the prediction. If a topic is pushing the decision higher (line 9), then its normalized value (i.e. its contribution) is added to c_sum^{+} (line 10) and its name is added to c_topics^{+} (line 12). Additionally, if the topic's normalized value is equal to or greater than the Opposing's bound, the topic is appended to the o_topics^{+} set (lines $12-13$). In contrast, if a topic is instead pushing the decision lower (line 14), its value is added to c_sum^{-} (line 15) and its name is appended to the c_topics^{-} set (line 16). Moreover, if the topic's normalized value is equal to or greater than the Opposing's bound (line 17), the topic is appended to the o_topics^{-} list (line 18). Once all topics of a given image have been cycled through and the variables have been assigned values, the algorithm is able to assign the image to a category (i.e. Collaborative or Opposing) based on its features.

If the image has at least one topic pushing the decision *sufficiently* higher (i.e. o_topics^{+} contains a topic) *and* at least one topic pushing the decision *sufficiently* lower (i.e. o_topics^{-} contains a topic), the image is categorized as Opposing (line 19) and the algorithm outputs the appropriate explanation using the most important topics and topic tags (lines 20-22).

If the image is not in the Opposing category, the algorithm instead checks if its topics are acting in a collaborative manner. It first checks whether the sum of all contributions pushing the decision higher (c_sum^{+}) is equal to or greater than the Collaborative's bound (line 23). If true, it means there are *enough* topics pushing the decision *decisively* higher. The top three most contributing topics (c_topics^{+}) and associated tags are subsequently returned (lines $24-25$) and the explanation is

outputted. On the other hand, if the image's topics are not collaboratively pushing the decision higher, the algorithm checks whether the opposite is true. Hence, if topics are together pushing the decision *decisively* lower and c_sum^{-} is thus equal to or greater than the Collaborative's bound, the image is considered Collaborative and the appropriate explanation is generated (lines $26 - 28$).

Finally, if the image does not belong to any of these categories, the algorithm considers the image as Weak and an explanation featuring the most relevant three topics and associated tags is generated (lines $29 - 32$). The algorithm ultimately generates as output an explanation of why the image has been classified as public or private. PEAK's implementation is available at `https://github.com/aycignl/peak`.

4.3. Evaluation

We evaluate the performance of our proposed system in terms of its contribution to preserving privacy using explanations.

User Study: We first aim to answer the following research question:

RQ2.1 Are the generated explanations by PEAK sufficient, satisfying, and understandable for humans?

We perform an online user study to evaluate our proposed explanation model in terms of sufficiency, satisfaction, and understanding. We conduct a pilot study (with $n = 5$ participants) before the real study to test whether the study is understandable. Based on the comments during the pilot, we improved the initial description of the study and reworded one question.

Our user study has three phases. In the first phase, we present a plain language statement that describes the study and a consent form. The second phase is meant to explain the study over an example, wherein we show an image, its generated expla-

nation, and the three questions that will be asked to the participant. Finally, in the third phase, each participant is exposed to 16 images with generated explanations in random order. Two of these images deliberately provide irrelevant explanations so that we can differentiate the participants that are attentive during the survey. Thus, these questions are meant to filter out the participants who are not focused. Such users are removed from the analysis.

In order to examine our explanation model, we personalize the *Explanation Satisfaction Scale* proposed by Hoffman *et al.* [77]. We ask participants to rank the following questions:

(i) This explanation that the algorithm produces has SUFFICIENT DETAIL.
(ii) This explanation produced by the algorithm is SATISFYING.
(iii) From this explanation, I UNDERSTAND why an image has been identified as private or public.

Each factor is accompanied by a $5-$point Likert scale (Strongly agree $= 5$, Somewhat agree $= 4$, Neither agree nor disagree $= 3$, Somewhat disagree $= 2$, Strongly disagree $= 1$). In the final phase, participants responded to anonymously collected demographic questions (age, gender, and education level) and optionally provided free-form text for comments/feedback. We designed our user study using the Qualtrics online survey tool[9].

Participants: A total of 57 participants responded to questions but we excluded 12 of them who did not answer the check questions properly. 64% of the remaining 45 participants were male and 36% were female. 26 participants were between $25-34$ years old, 14 were between $18-24$, 4 were between $35-44$, and 1 was between $55-64$. In terms of the highest degree of education, 19 of them had a Master's degree, 11 of them had a Bachelor's degree, six of them were High school graduates, five of them attended Some college ($1-4$ years, no degree), two of them had Doctorate degree, and

[9]https://www.qualtrics.com

two of them had Professional school degree (MD, DDC, JD, etc). This demographic is well-balanced on gender, age group, and education.

Results: We show our results of the user study in Figures 4.7, 4.8, 4.9, and 4.10, starting with the results for all images, then looking at the public and private classes specifically, and finally discussing the results for different explanation categories.

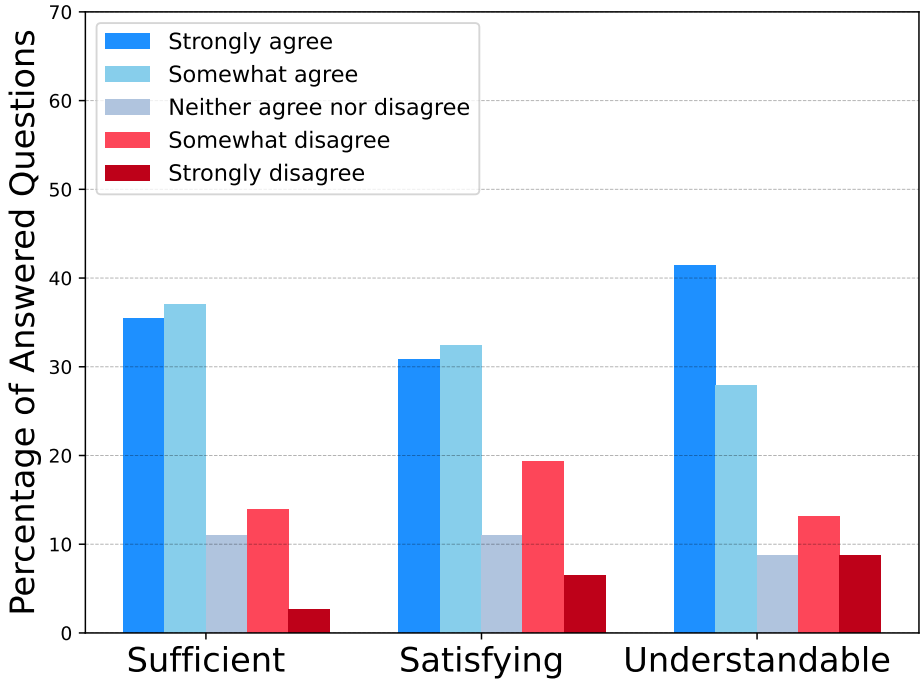


Figure 4.7. The answers for all questions.

Figure 4.7 shows participants' responses for all images for the three dimensions (i.e. Sufficient, Satisfying, and Understandable). First, the explanations were deemed to be sufficiently detailed ($M = 3.88, SD = 1.12$) as over 70% of respondents answered they either "strongly agree" or "somewhat agree", with the result statistically significant (p-value < 0.05). Moreover, only a low proportion of respondents indicated they "strongly disagree".

Second, the explanations were generally seen as satisfactory ($M = 3.62, SD =$

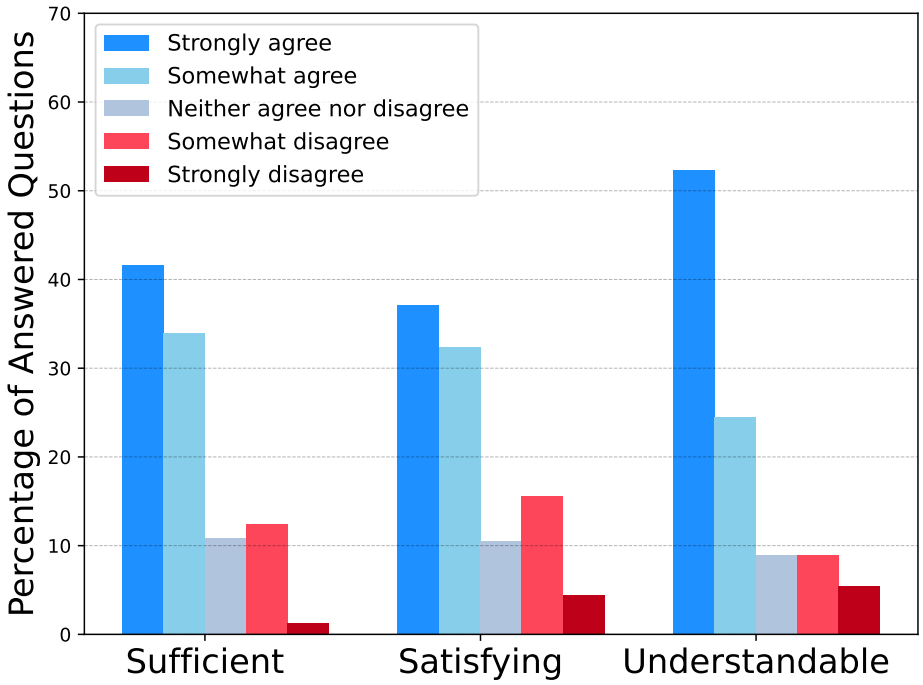


Figure 4.8. The answers for the questions with respect to the public class.

1.28) by respondents, a statistically significant result at the 0.05 significance level. More than 60% of participants answered in agreement and whilst there was some disagreement reported, this was largely only "somewhat disagree".

Finally, the explanations for why images were labeled as public or private were generally deemed to be understandable by participants ($M = 3.8, SD = 1.33$), again statistically significant at the 0.05 level. Out of all three dimensions, "strongly agree" was the most commonly given answer at over 40% of respondents, followed by close to 30% of "somewhat agree" answers. The range of answers was more varied however, as indicated by the larger standard deviation and a higher proportion of "strongly disagree" compared to *Sufficient* and *Satisfying*. Participants generally agreed the explanations were sufficiently detailed, they found the explanations satisfactory, and they understood why the images were labeled as private or public.

Figure 4.8 and 4.9 show the distributions of answers for the survey questions with

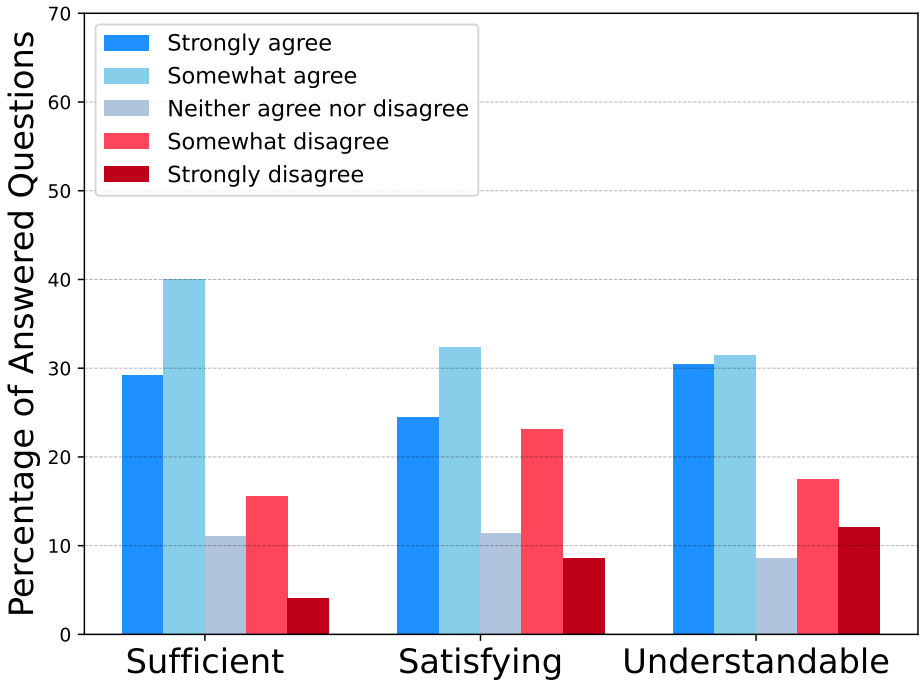
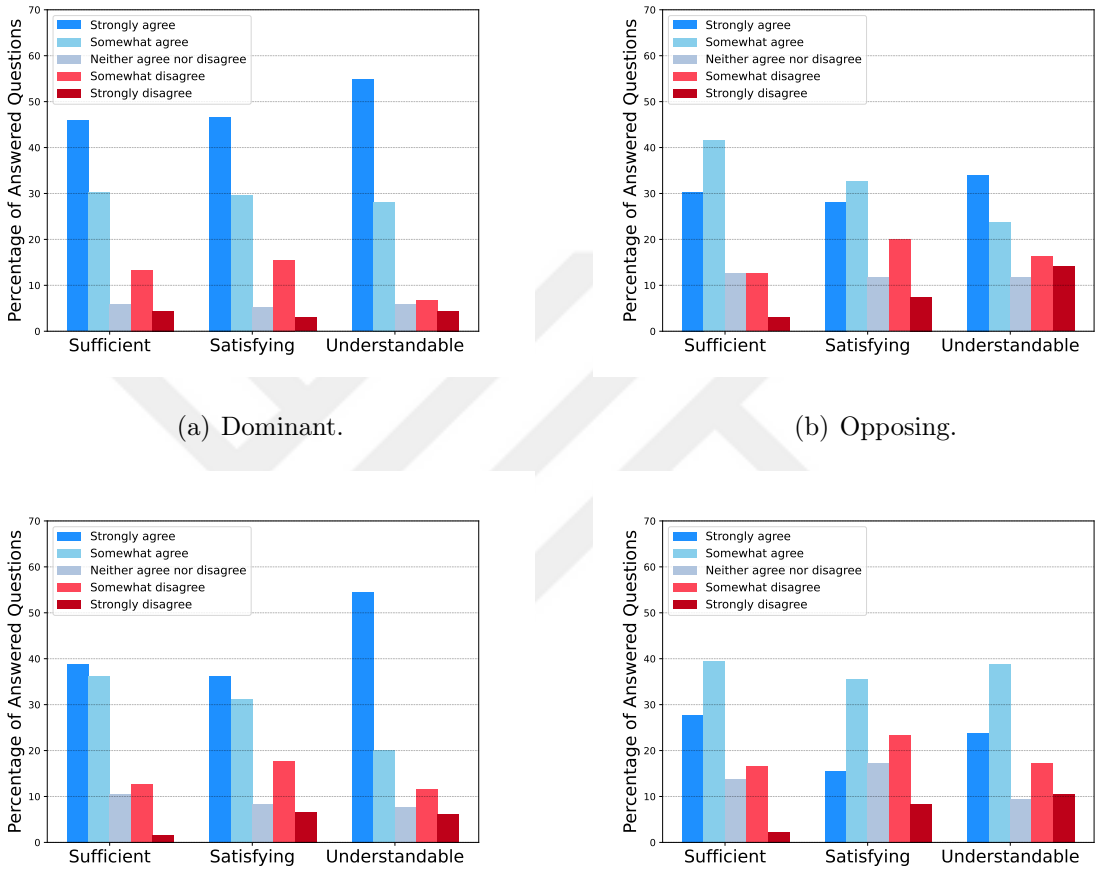


Figure 4.9. The answers for the questions with respect to the private class.

respect to the public and private classes. Figure 4.8 indicates that participants found the explanations for public images to be more sufficient, more satisfying, and more understandable compared to the images labeled as private. For all three dimensions, participants agreed ("strongly agree" and "somewhat agree" combined) more frequently for public images as opposed to private images. Additionally, "strongly agree" was the most frequent answer for public images in contrast to private images where "somewhat agree" is the most commonly given answer. For private images, there was more disagreement reported, with higher shares of "somewhat disagree" and "strongly disagree". For sufficiently detailed and satisfying, the differences between public and private images are not that large but for understandable there is quite a stark difference. For public images, more than half of respondents strongly agreed the explanations were understandable whereas for private images the share is only around 30% and nearly equal to the number of "somewhat agree" answers. Moreover, more respondents disagreed, and quite strongly so, for private images. A possible explanation for why participants generally answered "strongly agree" more often for public images across all three di-

mensions may be because privacy and what are considered private images is subjective and more difficult to explain. Hence, respondents agree less often and their responses vary more for private images compared to public images.



(a) Dominant.

(b) Opposing.

(c) Collaborative.

(d) Weak.

Figure 4.10. The answers for the questions with respect to the categories.

Figure 4.10 shows the distributions of answers to assess sufficiency, satisfaction, and understandability with respect to different categories (i.e. Dominant, Opposing, Collaborative, and Weak). Figure 4.10(a) and 4.10(c) demonstrate when an explanation has a decisive topic or is composed of like-minded topics in the decision, participants agree that the explanations of images belonging to such categories are sufficiently detailed and satisfying, with around 75% of participants either somewhat or strongly agreeing. Additionally, participants understand why an image is identified as belonging to a certain class (private or public) with more than half of participants strongly

agreeing and between 20% and 30% somewhat agreeing. On the other hand, compared to the Dominant and Collaborative categories, Figure 4.10(b) shows that participants score understanding ($M = 3.47, SD = 1.45$) of the decision less when an explanation has topics that have opposing forces in the decision. The images in this category have Opposing topics in terms of the contribution to the decisions. Thus, making a decision is not straightforward for the images belonging to the Opposing explanation category when compared to the Dominant and Collaborative categories. Moreover, Figure 4.10(d) shows the results for the explanations of the images belonging to the Weak category. For this category, the dispersion of responses is greater and it has the fewest number of participants answering strongly agree on all dimensions. This is not surprising as images in this category don't have easily identifiable characteristics like the other categories. Even if participants are only moderately confident ($M = 3.27, SD = 1.22$) that the explanations are satisfying, they do agree on the sufficiency of the explanations and understandability of a class decision based on the explanations. Overall, whilst there is some variation between different classes and image categories, participants generally find the explanations generated by PEAK as sufficient, satisfying, and understandable. In light of these results from the user study, we answer RQ2.1 as participants find generated explanations by PEAK sufficient, satisfying, and understandable.

Enhancing Privacy Assistants: We also evaluate the performance of PEAK in terms of enhancing the privacy assistant PURE by helping them make predictions about difficult-to-classify images through the use of image explanation categories. Moreover, PEAK can help reduce the number of images having to be delegated to the user for a decision. We aim to answer the following research question:

> *RQ2.2* Can the generated explanations of PEAK be used by personal assistant PURE to improve its decision-making?

PEAK can help PURE to classify images that are difficult to make a prediction about, e.g. low confidence predictions. It does so by dividing images into explanation

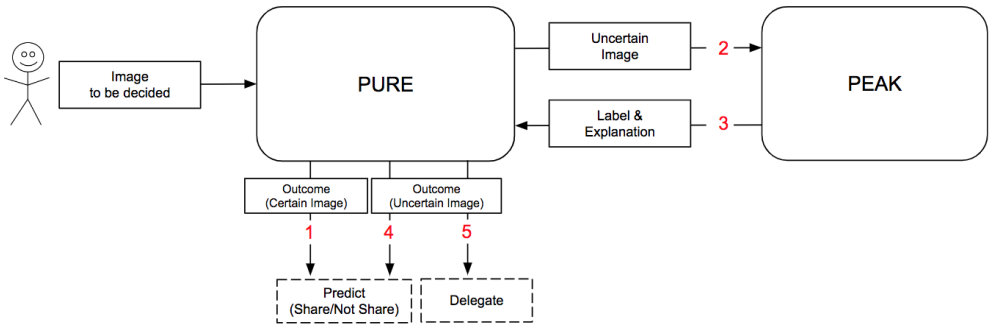


Figure 4.11. System overview schema of PEAK in combination with PURE.

categories, which allows it to uncover distinct and hidden relationships between topics for each category. The explanation categories are subsequently used to generate explanations for why an image is considered public or private.

PEAK is used in conjunction with personal assistant PURE, which is detailed in Chapter 2. Whilst PURE is a useful approach because no predictions are made for uncertain images, thus resulting in fewer mistakes, it does mean the user themselves have to decide on those uncertain images. Ideally, the user has to manually label as few images as possible. PEAK allows privacy assistants to classify images accurately, even uncertain ones, ultimately resulting in fewer images having to be delegated to the user. The reduction in delegated images can help to ensure sufficient cognitive resources are available for the remaining decision-making left up to the user. Ultimately, the privacy assistants' decision-making will have been improved without compromising its performance. In order to reduce the number of images delegated to the user to decide on, PEAK is utilized for uncertain images, i.e. images PURE has difficulty making a prediction about. PURE only uses PEAK's privacy decisions for certain class-category pairs for which PEAK performs well. For images where PEAK does not perform well, PURE delegates the decision back to the user. Hence, PEAK was not able to provide certainty for that images. Figure 4.11 shows the schema of PEAK in combination with PURE. A user gives an image to her personal privacy assistant PURE, which then reports its privacy prediction (i.e. share or not share) for the given image when certain about its prediction (1). When not certain, PURE first delegates the (uncertain) image to PEAK (2). In this case, PEAK classifies the uncertain image and subsequently

	Dominant		Opposing		Collaborative		Weak	
	public	**private**	**public**	**private**	**public**	**private**	**public**	**private**
All	5	8	9	6	30	32	5	4
Uncertain	3	11	13	12	11	37	7	6

Table 4.1. Percentage (%) of all and uncertain images in the Train-set that belong to the explanation categories.

shares the decision and the generated explanation with PURE (3). After receiving input from PEAK, PURE finalizes its decision whether to use PEAK's prediction result (i.e. share or not share) (4) or to delegate the privacy decision to the user (5).

As shown in Chapter 2, when the uncertainty threshold, denoted as θ, is set to 0.7 by a user, 66% of the images can be classified with PURE, i.e. these are certain images and accuracy of 96% is achieved [60]. The remaining 34% of images, a total of 1695, are all uncertain images and these are delegated to the user by PURE. Hence, if PEAK is able to improve on this percentage and reduce the number of images delegated to the user whilst at the same time maintaining similar accuracy, it mitigates the cognitive load in the user and successfully improves upon PURE.

Table 4.1 shows distributions of all and uncertain images in the Train-set across explanation categories. First, it is clear that the Collaborative category is the largest category since it comprises more than half of all images (62%) whereas the other categories are featured in roughly similar frequencies. In terms of privacy class, there is no clear bias towards either public or private images with the difference at most 3% within each category. Looking at only uncertain images, the Collaborative category is again the most common at 48% of uncertain images, but there are much fewer public images compared to all images (11% versus 30%) and somewhat more private images. A second difference for uncertain images is that there are more images in the Opposing category (25% compared to 15%). For the Dominant and Weak categories, there are only limited differences between all and only uncertain images. Finally, in contrast to all images, uncertain images more frequently belong to the private class (66% compared

to 50%).

	Dominant		Opposing		Collaborative		Weak	
	public	**private**	**public**	**private**	**public**	**private**	**public**	**private**
All	86	**88**	90	56	98	**93**	80	72
Uncertain	54	**86**	85	55	91	**91**	70	73

Table 4.2. Explanation category and class-specific privacy prediction performance (% accuracy) of PEAK on all and uncertain images in the Train-set.

PURE decides whether to use the output of PEAK to make predictions or to delegate decisions for uncertain images back to the user based on two criteria, namely *high* and *consistent* performance of PEAK. Hence, PEAK should perform well, i.e. accuracy should be greater than 0.85, and model performance on both *all* and *uncertain* images should be consistent, i.e. the difference in accuracy between the two groups should be less than 5%. Table 4.2 shows the privacy prediction performance of PEAK in terms of explanation category and class pairs. Based on the defined criteria, there are two *Category-class* pairs with high and consistent performance, namely *Dominant-private* and *Collaborative-private*. Hence, for uncertain images belonging to these two *Category-class* pairs, PEAK will be used. PEAK will share the label of the image and the generated explanation for the image specifying to which Category-class it belongs, along with the most important topics contributing to its decision.

The usefulness of PEAK should ultimately be determined based on performance for the uncertain images in the Test data. In the Test-set , Dominant-private and Collaborative-private images comprise 7% and 25% respectively, capturing nearly one-third of all uncertain images. Additionally, PEAK performs well as prediction accuracy is 94% and 93% for Dominant-private and Collaborative-private, respectively. Hence, high accuracy that's broadly in line with the observed performance for the Train-set. Overall, for nearly a third of the uncertain images, PEAK yields high performance. As for uncertain images belonging to the remaining *Category-class* pairs, PURE will not use PEAK's prediction as performance is not high and consistent for these pairs. Consequently, these uncertain images will be delegated to the user for a

privacy decision. We note, however, that this is the same outcome as when PEAK is not used, i.e. these are images which would have been delegated to the user regardless.

When PURE delegates the selected uncertain images to PEAK, fewer images are delegated to the user whilst classification accuracy is not compromised. Based on the Test-set, the number of images delegated to the user is 23%, whereas it is 34% for PURE alone. Hence, it mitigates the cognitive load of having to make a privacy decision for a potentially large number of images since user input is no longer needed for 542 uncertain images. Indeed, following the addition of PEAK, user input is only required for roughly one out of every four images as opposed to one out of every three when PURE is utilized. It reduces the risk of errors and biases by asking the user less, so a strong improvement due to PEAK. Moreover, given the high accuracy for Dominant-private and Collaborative-private images, overall model performance is high at 0.959, and nearly equal to PURE's standalone accuracy of 0.963. Consequently, combining PURE with PEAK results in nearly equal performance but for 77% of the data as opposed to 66% when it is only PURE.

Finally, it's worth adding that in general, making a prediction for private images is by itself already challenging. However, PEAK is able to make accurate predictions in an even more challenging situation, namely uncertain private images. In doing so, PEAK is effective in reducing the number of decisions delegated to the user whilst at the same time achieving convincing privacy prediction performance. Hence, in view of these results, we answer RQ2.2 positively, namely PEAK's generated explanations can indeed be used to improve personal assistants' decision-making.

4.4. Conclusion

In this chapter, we propose a novel privacy assistant PEAK to understand why a given image is considered public or private based on the characteristics of the image. Additionally, PEAK automatically generates explanations for its privacy decisions. Our privacy assistant is able to discover 20 latent topics from descriptive image tags using

topic modelling and subsequently makes privacy predictions based on the relationship between images and their associated topics. We restrict the number of topics to simplify explanations while keeping predictions as accurate as possible using these topics. Moreover, to further enhance ease of understanding for the user, we manually name each topic based on its associated tags, e.g. *Nature* for nature related tags. However, users are free to name topics themselves to suit their individual needs and to best understand PEAK's explanations. Each topic serves as a representation for a collection of visual contexts that are linked to its associated set of tags. This simple construct makes topics very flexible due to their limited semantic constraints, unlike a single textual representation of each topic. The approach works well as the privacy classifier achieves high accuracy, demonstrating the effectiveness of the topic-based representation of images. Furthermore, a user study shows the generated explanations are found to be sufficiently detailed, satisfying, and understandable, allowing us to positively answer RQ2.1. Finally, our results show that PEAK can improve the decision-making of personal assistants, thereby answering RQ2.2 affirmatively, by reducing the number of images delegated to the user whilst not compromising on model performance. In light of these results, we can positively answer Research Question 2. This chapter has been presented and published in [78–80].

5. LITERATURE SURVEY

The literature on approaches that help users manage their privacy is broad and spans comprehensive areas such as uncertainty and risk models, privacy prediction approaches, and explanation models. One of the earlier works is from Fang and LeFevre [81], who introduces a wizard software based on an active learning paradigm. The wizard generates a privacy preference model using extracted features from visible data and communities as well as user input such as asking questions. The wizard recommends privacy preferences to users for different information items on their profiles, such as birthday, address or telephone number. One of their key findings is that a user's social network structure is an important resource when modeling the user's privacy preferences. This idea has been exploited also by Kepez and Yolum [82], who propose an ML-based model for image privacy prediction. Their framework is based on several attributes about posts such as the sharing time, the location, and the content of the post. They utilize the user's social network to improve the predictions. Both of these approaches are important for the privacy prediction task because they use information about the user, her network, and her posts to improve predictions. However, in situations where such external information is not available, there is still a need for user recommendations that are solely based on the content itself.

Yu *et al.* [83] propose an algorithm to recommend privacy labels for images in OSNs. Their recommendation algorithm takes into account two approaches: the image content sensitiveness and users' trustworthiness. To train a tree classifier, the algorithm uses feature-based and object-based approaches for the image content sensitiveness and the characterization of users' trustworthiness based on social behaviors. Through extensive evaluations of a user study and two publicly available datasets (PicAlert and MIRFLICKR), they have shown the proposed algorithm is effective. Other work that takes into account trustworthiness in its approach is Kokciyan and Yolum [84]. They propose an approach, TURP, that manages the trustworthiness of information sources, IoT devices, for making context-based privacy decisions. They represent IoT devices

and users as software agents. Each agent has a confidence value when it shares information with another agent. In the beginning, each device has the same trust value. These values are updated based on feedback received from multiple agents. TURP uses Disjunctive Datalog while reasoning about information collected from multiple agents. It would be interesting to couple TURP with our proposed approach in an IoT context so that the privacy decisions are augmented with trust.

Han *et al.* [85] propose a method that uses multi-level and multi-scale deep representations for the image privacy prediction task. First, they obtain these deep representations from a CNN-based model. Then, they propose two feature aggregation models, Privacy-MSML and Privacy-MLMS, based on different aggregation strategies using Bi-LSTM and self-attention. They evaluate the performance of proposed models on a subset of the PicAlert dataset. They show that their proposed aggregation models yield better performance by F1-score compared with ResNet-18, CNN-RNN, and concatenated multi-level features. However, they are not concerned with capturing the uncertainty in their predictions as we have done here.

Jiao *et al.* [86] design a system, IEye, that provides a personalized and interpretable privacy model. They first extract features from images and use multi-layered semantic graphs for feature representations of images. Then they learn personalized privacy rule sets from images using the rule-based classification algorithm RIPPER. They compare their methods with SIFT and deep features extracted from pre-trained networks AlexNet, VGG16, and ResNet152. They evaluate the performance of their method IEye on the PicAlert dataset and a small dataset called PPP, which consists of 8744 images from 20 users. IEye's accuracy on the PPP dataset is higher than for the baseline approaches. However, the proposed method does not perform beteter than deep features for the PicAlert dataset.

Liu *et al.* [87] present problems, challenges, approaches, and future directions of ML in privacy. By taking into account the roles of ML in privacy, they divide existing works into three categories; private ML, ML enhanced privacy protection, and

ML-based privacy attacks. In the first category, the aim is to develop a private ML system including model parameters, Train and Test sets, and predictions. Differential privacy is one of the popular ML solutions capable of protecting the privacy of individual data items [88]. In the second category, ML approaches are used to make decisions for content with the aim to preserve privacy and predict information leakage. For instance, ML-based models predict applications' privacy risks, identify contents' sensitive information, and learn users' privacy preferences. The third category presents the importance of ML attack models, including re-identification and inference attacks. Preserving privacy under such attacks becomes more challenging with the rapid increase in the usage of online social networks and the recent advances in deep neural networks (DNNs). However, DNNs are vulnerable to adversarial perturbation. Goodfellow *et al.* [89] propose the Fast Gradient Sign method to generate adversarial examples. Differently from such kind of attacks, Miao *et al.* [90] introduce the controlled (protected) information stealing attack. They explain the phases of ML-based attack methodology, discuss the challenges of such attacks, and share the defense mechanisms. The work that we present here falls into the second category, namely that we use ML methods to enhance privacy protections.

5.1. Uncertainty and Risk Models

Sensoy *et al.* [47] propose a novel model to quantify the predictive uncertainty of classification tasks using EDL. Their model is designed to allow for the representation of uncertainty in the form of probability distributions, which are propagated through the network during training. EDL is based on Dempster-Shafer's theory of evidence and SL to quantify uncertainty in classification tasks. EDL helps to improve the reliability of DL models by quantifying uncertainty effectively. Sensoy *et al.* [46] also propose risk-calibrated evidential deep classifiers improve classifications by decreasing the costs of misclassified predictions. They reformulate the EDL model in order to accomplish this goal. Their experiments show that the proposed Risk EDL method has lower misclassification costs compared to EDL, standard learning with cross-entropy loss, and cost-sensitive learning methods for MNIST, FashionMNIST, and CIFAR10

datasets. They also report that their method is robust for out-of-distribution samples.

Collu *et al.* [91] present an access control mechanism to evaluate privacy risks for personalized content. It is important to notice that in the proposed scheme, users need to manage a single social network and people interact with their public keys securely with other people in their social networks. Carminati *et al.* [92] also propose an access control mechanism for OSNs. Their approach is rule-based for specifying access control policies by content owners in OSNs.

Akcora *et al.* [93] propose a risk measure for OSNs in terms of disclosure of personal information. For risk estimation, they use active learning approach. While estimating risk levels, they take into account second-level contact as *strangers*, the risk attitude of users to ask for owner feedback and risk judgment. Their proposed risk learning process is based on supervised learning techniques [94]. They test the proposed approach on real Facebook data and their approach is able to correctly predict risk labels with 83.38% of accuracy. However, they use only 47 Facebook users during the test who are from 5 different countries. Similarly, Akcora *et al.* [95] focus on finding the risks of friends caused by their network in OSNs. They find that having friends from different friend clusters with a friend of a friend has a big impact on the risk perception of users.

Yeom *et al.* [96] investigate the relationship between overfitting and privacy risk in ML models. They propose several regularization techniques to mitigate the privacy risk associated with overfitting. These techniques encourage simpler models and reduce the amount of noise memorized by the model. They also suggest using privacy-preserving mechanisms, such as differential privacy, to prevent the model from memorizing sensitive information. They show that mitigating overfitting can significantly reduce the privacy risk in ML models.

5.2. Privacy Prediction Approaches

5.2.1. Tag-based Approaches

An an alternative to approaches that rely on image content, some recent methods have used tags associated with content to predict privacy labels of images. Squicciarini *et al.* [97] introduce the Tag-To-Protect (T2P) system that automatically recommends privacy policies using the image's tags and the privacy policies. Their proposed system is useful for both newly uploaded images and cold-start problems when there are very few tags available. One of the prominent results from their experiment is that the prediction accuracy decreases when there is a large set of tags as a higher number of tags makes finding a pattern more difficult. Kurtan and Yolum [33] propose an agent-based approach predicting the binary privacy labels of images, i.e. private or public, using automatically generated content tags. The system keeps track of content being shared using tag tables. The internal tag table stores the data of privacy labels collected from images the user shares herself. The external tag table stores the data collected from the images the user's friends have shared with the user. Using metrics inspired by information retrieval, they define metrics to measure how informative a tag is to assess the privacy of an image. Contrary to previous approaches, this system performs well even when the personal assistant only has access to a small amount of data. However, they are not concerned about capturing the uncertainty explicitly or taking into account personal risk factors as we have done with PURE.

5.2.2. Textual- and Visual-based (Hybrid) Approaches

Various approaches have utilized textual and visual features to train classifiers. Earlier work is by Zerr *et al.* [31], where they propose a learning model for estimating privacy settings of images with binary labels (i.e. public or private). Their algorithm obtains the best performance for a hybrid combination of visual and textual information. However, they do not consider personalization, uncertainty, or explanation of the prediction. Tran *et al.* [98] propose a privacy framework, called Privacy-CNH

that consists of object and convolutional features using a CNN for image privacy detection. Similarly, Squicciarini *et al.* [99] present a learning model for binary privacy labels of images as well as multi-class privacy labels such as *Only You* or *Family*. They show that combining SIFT and tag features performs better than the other two or three combinations such as sentiment, RGB, or facial detection. The results of these approaches have been improved by Tonge and Caragea [32], who tackle the same problem using deep visual semantics (i.e. deep tags) and textual features (i.e. user tags) to develop a model that predicts whether an image is private or public. Deep tags of images are the top k predicted object categories extracted from pre-trained models. Utilizing user-created tags, they create deep visual features by adding highly correlated tags to visual features extracted from the fully connected layer of the pre-trained models. They use Support Vector Machine (SVM) classifiers with pre-trained CNN architectures such as AlexNet, GoogLeNet, VGG-16, and ResNet to extract features (tags). They find that a combination of user tags and deep tags from ResNet with the top 350 correlated tags performs well for privacy prediction, as adding the highly correlated tags improves prediction performance. While their focus has been on classification alone, we take into account both the uncertainty and the personalization associated with making privacy decisions, which is critically important for real-life use cases.

5.2.3. Policy-based Approaches

Squicciarini *et al.* [100] propose an Adaptive Privacy Policy Prediction (A3P) system that predicts a privacy policy for images based on the information available for a given user in the context of social networks. A3P needs a user to specify some privacy policies before making a prediction of privacy policies. When recommending a privacy policy, A3P takes into account significant resources for the privacy concept such as actual image content, metadata, and social circle. A3P consists of two main components: A3P-core and A3P-social. When a user uploads an image, the A3P-core classifies the image first based on its contents and then, updates each category into subcategories based on its metadata (if available). Then, A3P-core either predicts a policy based on historical behavior or invokes A3P-social. A3P-social finds representative privacy

policies using the user's social circle.

Fogues *et al.* [23] present a personal agent, SoSharP that recommends sharing policies in multi-user scenarios. SoSharP uses contextual-based, user-based, preference-based, and group-based features. These features help to provide personalized recommendations in three rounds. SoSharP makes recommendations to each user by using context-based and user-based features in the first round. It moves to the second round if at least one user has not accepted the sharing policy. It uses preference-based features in addition to the features used in the first round. In the final round, it makes a recommendation for all users by using group-based features. Finally, if most of the users do not agree with the recommendation from the last round, SoSharP recommends manual resolution. Mosca and Such [101] also propose an agent, ELVIRA, which for multi-user settings benefits from recommending individual decisions to each user. While we do not consider multi-user settings in this dissertation, our work can be applied in multi-user settings to recommend privacy labels for each user before a group decision is taken.

Albertini *et al.* [102] propose a method to learn the habits of users and suggest customized access control policies to every user in an OSN. They formalize the concept of users' privacy preferences and describe a procedure that helps to convert the information gathered by association rules into access control policies. There are different studies the about Relationship-Based Access Control (ReBAC) paradigm [103, 104] to control information sharing in OSNs. Carminati *et al.* [103] present an access control model for OSNs, where policies are specified in terms of the type, depth, and trust level of existing relationships. Fong *et al.* [104] develop a policy language to compose access control policies. Their model bases authorization decisions on the relationships between the resource owner and the resource accessor in OSNs.

5.2.4. Taxonomy-based Approaches

Taxonomy categorizes content and user preferences based on certain characteristics and attributes. It identifies important features and patterns in contents. Yuan

et al. [105] propose a context-dependent and privacy-aware model for images. The model uses the image's content and contextual information about the image, and a specific requester. Their proposed framework first extracts general features (e.g. people, location, time, activity) and contextual features of the sender's images. The system collects information about the sender's preferences by asking questions to the sender in different scenarios. It trains a classifier on this information. When a requester exists, the classifier makes a decision based on the sender's information and the requester's contextual features. They then evaluate their approach by conducting a user study on manually annotated images with personalized contextual sharing decisions through the *ProShare S* application that they developed. Orekondy *et al.* [106] present a model for the privacy risk prediction task for images and provide 68 privacy attributes such as *nudity, passport* and *religion*. Li *et al.* [107] propose a method to find out what kind of visual content is private. They develop a taxonomy with 28 categories such as *nudity/sexual, irresponsible to child* and *bad characters/unlawful/criminal*. Zhao *et al* [108] define a privacy taxonomy with 10 categories of the most commonly used descriptive keywords for a certain category. For example, the descriptive keywords of the category *religion/culture* include *culture, religion, and spiritual*. Even though they propose inspiring taxonomies for privacy, their approaches do not focus on uncertainty and do not provide explanations for a particular image as to why the image is considered public or private, as we have done in our work.

5.2.5. Group-based Approaches

An alternative set of approaches make use of groups of users, considering various similar aspects among users to make recommendations. Misra and Such [44] develop PACMAN, a personal assistant that recommends access control decisions. Their approach is based on identifying communities (such as friend networks) from the OSN structure of a user and information about the content, which is gathered through users being asked to manually select tags for the content. Zhong *et al.* [55] propose a Group-Based Personalized Model (GBPM) for an image privacy classification task. Their proposed model learns privacy groups and private content types. Using additional profile

information (e.g. gender or age), they estimate new users' privacy decisions. They evaluate their proposed model on a randomly selected subset of the PicAlert dataset [31] by first extending it with demographic and social network usage information. They show that GBPM (with profile information) outperforms several baselines such as SVM approaches.

5.3. Explanation Models

Gilpin *et al.* [109] provide a broad perspective of XAI systems and identify three categories of approaches for providing explanations in ML systems: processing, representation, and explanation-producing. Processing involves emulating the data processing of an ML system to establish connections between input and output in order to justify emitted choices. Representation refers to how the internal data structures and operations of the network can be explained to gain insight into why confident choices are made. Explanation-producing refers to generating human-understandable descriptions of the model's behavior and decision-making process. They suggest evaluating explanation-producing models by how well they align with human expectations. The role of human evaluation in interpreting ML models is significant as it allows for assessing the reasonableness of the explanations provided by the models. Riveiro and Thill [110] discuss the need for human evaluations in the field of XAI, specifically about how users interact with systems that generate explanations of AI systems. They highlight a gap between user expectations and the explanations provided by the system and suggest the need for more user-centered approaches to designing explanations. A factual explanation provides information about the reasons why a certain output was produced by an AI system, whereas a counterfactual explanation suggests alternative outcomes that could have occurred if the input to the system had been different. They find that factual explanations are appropriate when the system output aligns with user expectations. However, when there is a mismatch between expectations and output, neither factual nor counterfactual explanations are appropriate and counterfactual explanations may not be sufficient on their own. The results also suggest that the accuracy of the user's mental model of the AI system is connected to the effec-

tiveness of the explanations provided and that further exploration of the context and details of the global system model may be useful in this regard. Langley [111] explores the concept of explainable agency and its relationship with a normative and justified agency. The paper emphasizes the importance of creating intelligent agents that are capable of explaining their actions to humans in a way that aligns with human values and norms. They suggest that this can be achieved by designing agents that are able to reason about their own decisions and the reasons behind them, and by providing explanations that are understandable and relevant to humans. The paper proposes a framework for developing such agents, which involves incorporating normative principles into the agent's decision-making process and using explanations to justify the agent's actions. They highlight the importance of creating agents that are transparent, understandable, and aligned with human values in order to ensure that humans can trust and rely on them. Miller [112] examines studies of explainability within the scope of philosophy, social and cognitive psychology, and cognitive science. Their study provides various definitions of explainability, criteria for selecting explanations, evaluating explanations, and useful insights for XAI. They define the interpretability of a model as the degree to which the cause of a prediction can be understood. In the context of AI, generating an explanation that establishes a shared understanding of the decision-making process between an intelligent agent and a human observer is crucial. Justification is provided by explaining why a decision is good. Arrieta *et al.* [113] provide an overview of XAI by defining interpretability and explainability. They define interpretability as the ability to explain meaning in a form that people can understand. They associate explainability with an explanation as the interface between a human and a decision-maker.

The utilization of interpretable models enhances the degree of interpretability and transparency in their decision-making processes. Linear Regression, Logistic Regression, Decision Trees, Random Forest, and k-nearest neighbors provide interpretable insights into their predictions. For instance, the Random Forest model can be trained to predict whether a given image is public or private. It enables us to trace and visualize the decision paths to understand which features of the image had a major impact on

privacy predictions. While this is important since there can be thousands of features, it is not always possible to understand the explanations.

5.3.1. Model-specific Methods

These explanation methods improve our understanding of how deep neural networks make decisions. The most popular model-specific techniques include the Saliency Map and Attention Map. Simonyan *et al.* [36] propose methods for generating visualizations of CNNs classification models trained on the ImageNet dataset by utilizing numerical optimization of the input image. Even though their proposed saliency map of a given image helps to identify the regions of the image that are most discriminative with respect to the given class, it can be limited in capturing all the important information that flows through a deep neural network. Selvaraju *et al.* [114] present Gradient-weighted Class Activation Mapping (Grad-CAM) method that generates visual explanations for any CNN, regardless of its architecture. As a visual explanation for a given image, Grad-CAM provides a heatmap that would have the greatest impact on the output. Xiao *et al.* [115] propose a two-level attention model for a fine-grained classification task that can highlight discriminative regions at different levels of abstraction in an image. The first level of attention involves object-level filtering that selects relevant patches to feed into the classifier. The second level of attention conducts part-level detection to detect the parts of the classification. Since these methods are highly dependent on their architectures, their explanations cannot generalize well to different architectures (i.e. not CNN). Additionally, the units of attention-based approaches are not explicitly trained to provide human-understandable explanations. They can be also vulnerable to adversarial examples, which have the potential to deceive a model and flip the prediction (i.e. misclassification) by doing so.

5.3.2. Model-agnostic Methods

These explanation techniques can be applied to any ML model, regardless of its architecture. Model-agnostic methods for the binary classification task consider what

features of the input have been influential on the decision. Ribeiro *et al.* propose [35] the local interpretable model-agnostic explanation model (LIME) that generates a new dataset consisting of perturbed samples and the corresponding predictions of the black box model. It then trains an interpretable model (e.g. Random Forest) on this new dataset. Lundberg *et al.* [37] propose a model-agnostic feature relevance explanation model (SHAP), that is based on a game theoretically Shapley values [116]. This method computes the contribution of each feature to the prediction output. Slack *et al.* [117] show that LIME is more prone to being deceived compared to SHAP. LIME also may not be an appropriate method for explaining image privacy predictions, as it may not be able to effectively handle the complex relationships between features, where small changes can have a big impact on the model's predictions. Furthermore, the method of perturbing the input used by LIME can be easily recognized by attackers. On the other hand, SHAP explanations are based on Shapley values, which can be difficult to interpret by non-expert users due to the highlighting of the contributions of each feature. In addition, explanations may be more difficult to understand when SHAP has a high-dimensional feature space.

5.3.3. Example-based Explanations

These explanations utilize specific examples from the dataset to understand how minor changes in a given image would affect the model's prediction and how the model made the prediction for the image. Example-based explanations provide examples from the dataset instead of displaying feature importance. Wachter *et al.* [39] propose counterfactual explanations that are a way of understanding the causal relationship between inputs and the model's predictions by simulating hypothetical scenarios and analyzing how minor changes to the feature of a given image affect the prediction. The conventional notion of "explanation" in the AI literature involves explaining the internal state or logic of an algorithm that leads to a decision, whereas counterfactual explanations focus on the external factors that led to that decision. This is a critical distinction since ML algorithms can have millions of interrelated variables, making it difficult to convey their behavior to non-expert users. Counterfactual explanations help to interpret the

predictions; however, multiple counterfactual explanations can be generated for a given image. It can be challenging to decide which one to use in this case.

Despite the fact that these methods can assist in providing explanations for the model's predictions, they can also encounter difficulties that can make them unable to produce human-understandable explanations. For example, explanations can include too many features or they may not be representative of the model's behavior.

An important work of explainability in conjunction with privacy, is by Mosca and Such [34], where they develop an agent that uses computational argumentation to resolve disputes and propose a text-based description of the outcome of the argumentation, generated by the system. They suggest a framework for generating explanations consisting of two types of explanations: general and contrastive. A general explanation provides an overview of the agent's decision-making process, without providing specific details about the reasons behind a particular decision. They are useful for giving users an understanding of how the system works and what factors it considers when making decisions. On the other hand, a contrastive explanation provides specific reasons for a particular decision and is intended to help users understand how they might modify their behavior to achieve better outcomes. These types of explanations are particularly useful for situations where the agent's decision might conflict with a user's goals or preferences. While explaining how the system operates is useful, that work does not provide an explanation as to why a given piece of content is private or public. Hence, with PEAK our focus is on explaining to both end-users and personal assistants why a given image can be considered private or public. PEAK is a model-agnostic approach that can be applied to any ML model.

6. DISCUSSION

6.1. Summary of Contributions

In this dissertation, we have achieved our research objective of designing trustworthy privacy assistants to manage privacy. We have done so by building two distinct assistants. The first is PURE, an uncertainty-aware privacy assistant that helps users make privacy decisions whilst incorporating the ambiguity of privacy and users' privacy personas. The second is PEAK, a personal assistant capable of classifying images and generating explanations for why an image is considered private or public. Additionally, PEAK is able to successfully work with PURE an improve the latter's decision-making. Our work contributes to the literature in the following ways:

- PURE captures the ambiguity of privacy through its modeling of uncertainty and exploits it to refrain from making wrong privacy decisions for its users.
- PURE outperforms alternative predictive uncertainty models, with the improvement being statistically significant.
- PURE adjusts its behavior based on the personal risk and expectations of its user whilst less decisions are delegated to the user.
- We utilize textual features (i.e. tags and topics), which achieves state-of-the-art performance for image privacy prediction.
- PEAK generates sufficient, satisfying, and understandable explanations for humans, with participants in a user study reporting as such.
- PEAK can be used by PURE to improve decision-making by reducing the number of images delegated to the user whilst not compromising on model performance.

6.2. Threats to Validity

In this chapter, we explore potential threats to the validity of our research within the framework of Wohlin *et al.*'s [118] four categories: *conclusion validity*, *internal validity*, *construct validity*, and *external validity*. We can categorize and address these threats as follows:

Internal validity: Oversampling was implemented to balance class sizes, particularly in the private class. However, this approach may distort the classifier's learning process and potentially lead to increased misclassifications. The skewed distribution due to oversampling might affect classifier performance. Despite this potential impact, comparable results between the training and test samples suggest any effect of oversampling may be limited.

Construct validity: We used a subset of the publicly available PicAlert dataset. In this dataset, images are annotated by a different number of annotators. We assume an image is private if at least one annotator annotated it as private. It is public if all annotators are annotated as public. However, disagreement among annotators implies potential limitations in the representation of a true private image. This definition of privacy might not fully reflect the concept of private images, potentially influencing classifier performance. However, in our test sample we only include private images where all annotators agree, in contrast to the training sample where private images with conflicting labels are included. Since performance for the training and test samples are comparable, this suggests the potential impact of conflicting labels on performance is limited. Additionally, convincing recall results for the private class suggests relative robustness. Finally, since privacy is subjective, it may be unreasonable to expect complete agreement by annotators on all private images, and the presence of some conflicting labels could therefore simply be a reflection of the reality that privacy is ambiguous and subjective.

Our dataset consists of images labeled by external annotators rather than the

owners of the images. Utilizing external annotators rather than image owners might lead to disparities in labeling due to a potential loss in contextual information. For example, an image of a room might be labeled as public by an external annotator whilst the image owner might consider the image to be private because it shows personal belongings or other sensitive information. As a result, the performance of the model may not be fully representative of the performance using another dataset where the images are all labeled by the owners of the images. A potential mitigating factor is that only some images could be expected to have conflicting labels between external annotators and the image owner whilst for others there is no conflict and thus no impact.

External validity: Despite the demographic balance (i.e. gender, education, and age) in our user study, using a relatively limited number of private images (8) might not fully capture respondents' views on the explanations for the private class. Inadequate representation of private images could impact the assessment of the suitability of the generated explanations. Recognizing the subjective nature of privacy, incorporating more private images might be useful to get an even better understanding of respondents' views regarding the suitability of our explanations.

The absence of genuinely private images (e.g. an intimate image of someone) in the dataset, as individuals are very unlikely to share such images online, raises potential concerns about the model's applicability in classifying highly private images. The performance of the classifier might thus not be representative of how the model would perform in a setting involving the classification of such highly private images. To better assess this aspect, exploring a dataset containing users' most private images could yield insights into potential changes in the model's performance.

6.3. Future Directions

This work opens up interesting directions for future research. We currently started with an uncertainty-aware model for privacy classification and enhanced it further

with users' personalized risk for misclassification. An interesting direction for further research is to enable PURE to have deeper interactions with the user. For example, it could interact with the user to obtain labels for images it is uncertain about and further enhance its ability for classification utilizing this new personal data. Another important research direction is to enable interaction between different personal assistants to help create a collaborative environment for preserving privacy. Currently, we assume the content a personal assistant decides on belongs solely to the user. However, many times content, such as group images or co-edited documents, can belong to more than one user [30]. Extending PURE to act collaboratively in such settings would be useful. Another avenue for further research relates to the fact a user's privacy preferences are often relational. That is, a user might be fine with sharing content with a friend but not with a colleague. Our current approach does not capture with whom the content is shared. It would be interesting to learn the relation-based sharing behavior of users. Finally, it would be interesting to investigate how PURE can learn when to delegate its decisions to the user. Currently, PURE has a preset threshold value θ by capturing user involvement set by the user such that it makes a decision when a prediction's uncertainty value is below θ, and delegates the decision otherwise. It would be useful if PURE could automatically adjust θ based on user feedback and subsequently improve its decision-making.

We further build on PURE with PEAK, for which there are interesting research directions as well. An important direction for future work regarding PEAK is to be able to get feedback from users and update the explanations accordingly. For instance, after each explanation, participants might designate the first topic as useful, but express that the second topic was not, thereby indicating a chance to improve the user study questions and, ultimately, the explanations. In future work, we will deep dive into the Weak explanation category by clustering. For instance, we can discover groups in this category by utilizing k-means or hierarchical clustering algorithms [119,120]. Another possibility is to automatically generate tags using different models/tools, as well as Clarifai. For instance, multi-modal deep learning models such as CLIP [121] can predict which text snippets (or tags) are related to a given image (and vice versa) using distances in the

joint embedding space of text and images. Given public and private images, we can extract all relevant words in the proximity of the image embeddings and consider the most discriminative words as tags with tag diversity in mind. Moreover, instead of naming the generated topics manually, this could be done automatically. Various automated methods are available for this purpose, such as selecting the most similar word to the tags as the topic name. The similarity can be calculated using word embeddings of the tags or by utilizing an ontology like WordNet [122]. We leave the task of automatically naming the generated topics as future work. As an interesting line of future research, PEAK's explanation categories could be adapted to the Schwartz theory of basic values [123]. A final future direction would be expanding our methodology to the classification of confidential documents since our approach is flexible and can work with text-based inputs as well. These are all interesting topics that we defer to future work.

REFERENCES

1. Westin, A. F., *Privacy and Freedom*, Atheneum, New York, 1967.

2. Warren, S. and L. Brandeis, "The Right to Privacy", *Harward Law Review*, Vol. 4, pp. 193–220, 1890.

3. Glancy, D., "Getting Government Off the Backs of People: The Right of Privacy and Freedom of Expression in the Opinions of Justice William O. Douglas", *Santa Clara Law Review*, Vol. 21, p. 1047, 1981.

4. O'brien, D. M., *Privacy, Law, and Public Policy*, Praeger, New York, 1979.

5. Cavoukian, A., "Privacy by Design: The 7 Foundational Principles", *Information and Privacy Commissioner of Ontario, Canada*, Vol. 5, p. 12, 2009.

6. Sicari, S., A. Rizzardi, L. A. Grieco and A. Coen-Porisini, "Security, Privacy and Trust in Internet of Things: The Road Ahead", *Computer networks*, Vol. 76, pp. 146–164, 2015.

7. Schoeman, F. D., *Philosophical Dimensions of Privacy: An Anthology*, Cambridge University Press, Cambridge, 1984.

8. Such, J. M., A. Espinosa and A. García-Fornes, "A Survey of Privacy in Multi-agent Systems", *The Knowledge Engineering Review*, Vol. 29, No. 3, pp. 314–344, 2014.

9. Cook, P. and C. Heilmann, "Two Types of Self-Censorship: Public and Private", *Political studies*, Vol. 61, No. 1, pp. 178–196, 2013.

10. Hogben, G., "Security Issues and Recommendations for Online Social Networks", *ENISA position paper*, Vol. 1, pp. 1–36, 2007.

11. Fogues, R., J. M. Such, A. Espinosa and A. Garcia-Fornes, "Open Challenges in Relationship-Based Privacy Mechanisms for Social Network Services", *International Journal of Human-Computer Interaction*, Vol. 31, No. 5, pp. 350–370, 2015.

12. Madejski, M., M. Johnson and S. Bellovin, "The Failure of Online Social Network Privacy Settings", *Columbia University Computer Science Technical Reports, CUCS-010-11*, 2011.

13. Sleeper, M., R. Balebako, S. Das, A. L. McConahy, J. Wiese and L. F. Cranor, "The Post That Wasn't: Exploring Self-Censorship on Facebook", *Proceedings of the 2013 conference on Computer supported cooperative work*, San Antonio, Texas, USA, pp. 793–802, 2013.

14. Spiekermann, S., J. Grossklags and B. Berendt, "E-privacy in 2nd Generation E-commerce: Privacy Preferences versus Actual Behavior", *Proceedings of the 3rd ACM Conference on Electronic Commerce*, New York, NY, USA, pp. 38–47, 2001.

15. Iachello, G. and J. Hong, "End-User Privacy in Human-Computer Interaction", *Foundations and Trends in Human-Computer Interaction*, Vol. 1, No. 1, pp. 1–137, 2007.

16. Brown, B., "Studying the Internet Experience", *HP laboratories technical report HPL*, Vol. 49, 2001.

17. Barth, S. and M. D. De Jong, "The Privacy Paradox–Investigating Discrepancies Between Expressed Privacy Concerns and Actual Online Behavior–A Systematic Literature Review", *Telematics and informatics*, Vol. 34, No. 7, pp. 1038–1058, 2017.

18. Nissenbaum, H., "Privacy as Contextual Integrity", *Washington Law Review*, Vol. 79, pp. 101–139, 2004.

19. Nissenbaum, H., *Privacy in Context: Technology, Policy, and the Integrity of Social Life*, Stanford University Press, California, 2009.

20. Kökciyan, N. and P. Yolum, "Priguard: A Semantic Approach to Detect Privacy Violations in Online Social Networks", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 28, No. 10, pp. 2724–2737, 2016.

21. Hoofnagle, C. J., B. Van Der Sloot and F. Z. Borgesius, "The European Union General Data Protection Regulation: What It Is and What It Means", *Information & Communications Technology Law*, Vol. 28, No. 1, pp. 65–98, 2019.

22. Vila, T., R. Greenstadt and D. Molnar, "Why We Can't Be Bothered to Read Privacy Policies", *Proceedings of the 5th International Conference on Electronic Commerce*, New York, NY, USA, pp. 403–407, 2003.

23. Fogues, R. L., P. K. Murukannaiah, J. M. Such and M. P. Singh, "Sosharp: Recommending Sharing Policies in Multiuser Privacy Scenarios", *IEEE Internet Computing*, Vol. 21, No. 6, pp. 28–36, 2017.

24. Colnago, J., 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", *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, p. 1–13, 2020.

25. Myers, K., P. Berry, J. Blythe, K. Conley, M. Gervasio, D. L. McGuinness, D. Morley, A. Pfeffer, M. Pollack and M. Tambe, "An Intelligent Personal Assistant for Task and Time Management", *AI Magazine*, Vol. 28, No. 2, pp. 47–47, 2007.

26. Edu, J. S., J. M. Such and G. Suarez-Tangil, "Smart Home Personal Assistants: A Security and Privacy Review", *ACM Computing Surveys*, Vol. 53, p. 116, 2020.

27. Hauswald, J., L. Tang, J. Mars, M. Laurenzano, Y. Zhang, C. Li, A. Rovinski, A. Khurana, R. Dreslinski, T. Mudge and V. Petrucci, "Sirius: An Open

End-to-End Voice and Vision Personal Assistant and Its Implications for Future Warehouse Scale Computers", *ACM SIGPLAN Notices*, Vol. 50, pp. 223–238, 2015.

28. Kekulluoglu, D., N. Kokciyan and P. Yolum, "Preserving Privacy as Social Responsibility in Online Social Networks", *ACM Transactions on Internet Technology*, Vol. 18, No. 4, 2018.

29. Such, J. M. and M. Rovatsos, "Privacy Policy Negotiation in Social Media", *ACM Transactions on Autonomous and Adaptive Systems*, Vol. 11, No. 1, 2016.

30. Ulusoy, O. and P. Yolum, "PANOLA: A Personal Assistant for Supporting Users in Preserving Privacy", *ACM Transactions on Internet Technology*, Vol. 22, No. 1, 2021.

31. Zerr, S., S. Siersdorfer, J. Hare and E. Demidova, "Privacy-Aware Image Classification and Search", *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Portland, Oregon, USA, pp. 35–44, 2012.

32. Tonge, A. and C. Caragea, "Image Privacy Prediction Using Deep Neural Networks", *ACM Transactions on the Web (TWEB)*, Vol. 14, No. 2, pp. 1–32, 2020.

33. Kurtan, A. C. and P. Yolum, "Assisting Humans in Privacy Management: An Agent-Based Approach", *The 20th International Conference on Autonomous Agents and Multiagent Systems*, Vol. 35, Virtual Event, UK, pp. 1–33, 2021.

34. Mosca, F. and J. Such, "An Explainable Assistant for Multiuser Privacy", *Autonomous Agents and Multi-Agent Systems*, Vol. 36, No. 1, pp. 1–45, 2022.

35. Ribeiro, M. T., S. Singh and C. Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier", *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco,

CA, USA, pp. 1135–1144, 2016.

36. Simonyan, K., A. Vedaldi and A. Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps", *arXiv preprint arXiv:1312.6034*, 2013.

37. Lundberg, S. M. and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions", *Advances in Neural Information Processing Systems*, Vol. 30, 2017.

38. Lundberg, S. M., G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal and S.-I. Lee, "From Local Explanations to Global Understanding with Explainable AI for Trees", *Nature Machine Intelligence*, Vol. 2, No. 1, pp. 2522–5839, 2020.

39. Wachter, S., B. Mittelstadt and C. Russell, "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR", *Harvard Journal of Law & Technology*, Vol. 31, p. 841, 2017.

40. Adebayo, J., J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt and B. Kim, "Sanity Checks for Saliency Maps", *Advances in Neural Information Processing Systems*, Vol. 31, 2018.

41. Ghorbani, A., A. Abid and J. Zou, "Interpretation of Neural Networks is Fragile", *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, Hawaii, USA, pp. 3681–3688, 2019.

42. Kindermans, P. J., S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan and B. Kim, "The (Un)Reliability of Saliency Methods", *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pp. 267–280, 2019.

43. Szymanski, M., M. Millecamp and K. Verbert, "Visual, Textual or Hybrid: The Effect of User Expertise on Different Explanations", *26th International Conference on Intelligent User Interfaces*, New York, NY, USA, pp. 109–119, 2021.

44. Misra, G. and J. M. Such, "Pacman: Personal Agent for Access Control in Social Media", *IEEE Internet Computing*, Vol. 21, No. 6, pp. 18–26, 2017.

45. Acquisti, A. and J. Grossklags, "Privacy and Rationality in Individual Decision Making", *IEEE Security & Privacy*, Vol. 3, No. 1, pp. 26–33, 2005.

46. Sensoy, M., M. Saleki, S. Julier, R. Aydogan and J. Reid, "Misclassification Risk and Uncertainty Quantification in Deep Classifiers", *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2484–2492, 2021.

47. Sensoy, M., L. Kaplan and M. Kandemir, "Evidential Deep Learning to Quantify Classification Uncertainty", *Advances in Neural Information Processing Systems*, Montreal, Canada, pp. 3179–3189, 2018.

48. Dempster, A. P., "A Generalization of Bayesian Inference", R. R. Yager and L. Liu (Editors), *Classic Works of the Dempster-Shafer Theory of Belief Functions*, Vol. 219, pp. 73–104, Springer, Heidelberg, 2008.

49. Jøsang, A., *Subjective Logic: A Formalism for Reasoning Under Uncertainty*, Springer Publishing Company, Incorporated, 2016.

50. Zhang, J., M. Sensoy and R. Cohen, "A Detailed Comparison of Probabilistic Approaches for Coping with Unfair Ratings in Trust and Reputation Systems", L. Korba, S. Marsh and R. Safavi-Naini (Editors), *2008 Sixth Annual Conference on Privacy, Security and Trust*, Washington, DC, USA, pp. 189–200, 2008.

51. Kaplan, L., M. Şensoy and G. de Mel, "Trust Estimation and Fusion of Uncertain Information by Exploiting Consistency", *17th International Conference on Information Fusion (FUSION)*, Salamanca, Spain, pp. 1–8, 2014.

52. He, K., X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition", *Proceedings of the IEEE Conference on Computer Vision and Pattern*

Recognition, Las Vegas, NV, USA, pp. 770–778, 2016.

53. Szegedy, C., V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp. 2818–2826, 2016.

54. Simonyan, K. and A. Zisserman, "Very Deep Convolutional Networks for Large-scale Image Recognition", *arXiv preprint arXiv:1409.1556*, 2014.

55. Zhong, H., A. C. Squicciarini, D. J. Miller and C. Caragea, "A Group-Based Personalized Model for Image Privacy Classification and Labeling", C. Sierra (Editor), *The International Joint Conference on Artificial Intelligence (IJCAI)*, Vol. 17, Melbourne, Australia, pp. 3952–3958, 2017.

56. Gal, Y. and Z. Ghahramani, "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning", *The 33rd International Conference on Machine Learning (ICML)*, New York, NY, USA, pp. 1050–1059, 2016.

57. Lakshminarayanan, B., A. Pritzel and C. Blundell, "Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles", I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan and R. Garnett (Editors), *Advances in Neural Information Processing Systems*, Vol. 30, Long Beach, CA, USA, pp. 6402–6413, 2017.

58. Hüllermeier, E. and W. Waegeman, "Aleatoric and Epistemic Uncertainty in Machine Learning: A Tutorial Introduction", *arXiv preprint arXiv:1910.09457*, 2019.

59. Noreen, E. W., *Computer-intensive Methods for Testing Hypotheses*, Wiley, New York, 1989.

60. Ayci, G., M. Şensoy, A. Özgür and P. Yolum, "Uncertainty-Aware Personal Assistant for Making Personalized Privacy Decisions", *ACM Transactions on Internet*

Technology, Vol. 23, No. 1, pp. 1–24, 2023.

61. Ayci, G., "Uncertainty-aware Personal Assistant and Explanation Method for Privacy Decisions", *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, London, UK, pp. 2991–2992, 2023.

62. Lan, M., C. L. Tan and H.-B. Low, "Proposing a New Term Weighting Scheme for Text Categorization", *Association for the Advancement of Artificial Intelligence (AAAI)*, Vol. 6, Boston, Massachusetts, USA, pp. 763–768, 2006.

63. Devlin, J., M.-W. Chang, K. Lee and K. Toutanova, "Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding", *arXiv preprint arXiv:1810.04805*, 2018.

64. Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality", *Advances in Neural Information Processing Systems*, pp. 3111–3119, 2013.

65. Pennington, J., R. Socher and C. Manning, "Glove: Global Vectors for Word Representation", *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, pp. 1532–1543, 2014.

66. Breiman, L., "Random Forests", *Machine learning*, Vol. 45, No. 1, pp. 5–32, 2001.

67. Ho, T. K., "Random Decision Forests", *Proceedings of 3rd International Conference on Document Analysis and Recognition*, Vol. 1, Montreal, Canada, pp. 278–282, 1995.

68. Szegedy, C., W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich, "Going Deeper with Convolutions", *arXiv preprint arXiv:1409.4842*, 2014.

69. Krizhevsky, A., I. Sutskever and G. E. Hinton, "Imagenet Classification with Deep

Convolutional Neural Networks", *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.

70. Tonge, A. and C. Caragea, "On the Use of Deep Features for Online Image Sharing", P.-A. Champin, F. L. Gandon, M. Lalmas and P. G. Ipeirotis (Editors), *Companion Proceedings of the The Web Conference 2018*, Geneva, Switzerland, pp. 1317–1321, 2018.

71. Tonge, A. K. and C. Caragea, "Image Privacy Prediction Using Deep Features", D. Schuurmans and M. P. Wellman (Editors), *Thirtieth AAAI Conference on Artificial Intelligence*, Phoenix, Arizona USA, 2016.

72. Lee, D. D. and H. S. Seung, "Learning the Parts of Objects by Non-negative Matrix Factorization", *Nature*, Vol. 401, No. 6755, pp. 788–791, 1999.

73. Blei, D. M., A. Y. Ng and M. I. Jordan, "Latent Dirichlet Allocation", *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022, 2003.

74. Aycı, G., "Handling Uncertainty and Risk in Privacy Prediction", *1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020)*, Vol. 29, 2020.

75. Aycı, G. and P. Yolum, "Recommending Privacy Labels for Online content", *Proceedings of the 1st International Workshop on AI for Privacy*, European Conference on Artificial Intelligence (ECAI), 2020.

76. Sweller, J., "Cognitive Load During Problem Solving: Effects on Learning", *Cognitive science*, Vol. 12, No. 2, pp. 257–285, 1988.

77. Hoffman, R. R., S. T. Mueller, G. Klein and J. Litman, "Metrics for Explainable AI: Challenges and Prospects", *arXiv preprint arXiv:1812.04608*, 2018.

78. Aycı, G., A. Özgür, M. Şensoy and P. Yolum, "Can We Explain Privacy?", *IEEE*

Internet Computing, Vol. 27, No. 4, pp. 75–80, 2023.

79. Aycı, G., A. Özgür, M. Şensoy and P. Yolum, "Explain to Me: Towards Understanding Privacy Decisions", *The 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, London, UK, pp. 2790–2792, 2023.

80. Ayci, G., A. Özgür, M. Şensoy and P. Yolum, "PEAK: Explainable Privacy Assistant through Automated Knowledge Extraction", *arXiv preprint arXiv:2301.02079*, 2023.

81. Fang, L. and K. LeFevre, "Privacy Wizards for Social Networking Sites", *Proceedings of the 19th International Conference on World Wide Web*, North Carolina USA, pp. 351–360, 2010.

82. Kepez, B. and P. Yolum, "Learning Privacy Rules Cooperatively in Online Social Networks", *Proceedings of the 1st International Workshop on AI for Privacy and Security*, New York, NY, USA, p. 3, 2016.

83. Yu, J., Z. Kuang, B. Zhang, W. Zhang, D. Lin and J. Fan, "Leveraging Content Sensitiveness and User Trustworthiness to Recommend Fine-grained Privacy Settings for Social Image Sharing", *IEEE Transactions on Information Forensics and Security*, Vol. 13, No. 5, pp. 1317–1332, 2018.

84. Kökciyan, N. and P. Yolum, "TURP: Managing Trust for Regulating Privacy in Internet of Things", *IEEE Internet Computing*, Vol. 24, No. 6, pp. 9–16, 2020.

85. Han, Y., Y. Huang, L. Pan and Y. Zheng, "Learning Multi-level and Multi-scale Deep Representations for Privacy Image Classification", *Multimedia Tools and Applications*, pp. 1–16, 2021.

86. Jiao, R., L. Zhang and A. Li, "Ieye: Personalized Image Privacy Detection", *2020 6th International Conference on Big Data Computing and Communications (BIGCOM)*, Deqing, China, pp. 91–95, 2020.

87. Liu, B., M. Ding, S. Shaham, W. Rahayu, F. Farokhi and Z. Lin, "When Machine Learning Meets Privacy: A Survey and Outlook", *ACM Computing Surveys (CSUR)*, Vol. 54, No. 2, pp. 1–36, 2021.

88. Dwork, C., "Differential Privacy: A Survey of Results", *International Conference on Theory and Applications of Models of Computation*, Kraków, Poland, pp. 1–19, 2008.

89. Goodfellow, I. J., J. Shlens and C. Szegedy, "Explaining and Harnessing Adversarial Examples", *arXiv preprint arXiv:1412.6572*, 2014.

90. Miao, Y., C. Chen, L. Pan, Q.-L. Han, J. Zhang and Y. Xiang, "Machine Learning-Based Cyber Attacks Targeting on Controlled Information: A survey", *ACM Computing Surveys (CSUR)*, Vol. 54, No. 7, pp. 1–36, 2021.

91. Gollu, K. K., S. Saroiu and A. Wolman, "A Social Networking-Based Access Control Scheme for Personal Content", *Proceedings of the 21st ACM Symposium on Operating Systems Principles (SOSP'07)-Work-in-Progress Session*, New York, NY, USA, pp. 71–80, 2007.

92. Carminati, B., E. Ferrari and A. Perego, "Enforcing Access Control in Web-Based Social Networks", *ACM Transactions on Information and System Security (TISSEC)*, Vol. 13, No. 1, pp. 1–38, 2009.

93. Akcora, C., B. Carminati and E. Ferrari, "Privacy in Social Networks: How Risky is Your Social Graph?", *2012 IEEE 28th International Conference on Data Engineering*, Arlington, VA, USA, pp. 9–19, 2012.

94. Kotsiantis, S., "Supervised Machine Learning: A Review of Classification Techniques", *Informatica (Slovenia)*, Vol. 31, pp. 249–268, 2007.

95. Akcora, C. G., B. Carminati and E. Ferrari, "Risks of Friendships on Social Networks", *2012 IEEE 12th International Conference on Data Mining*, Brussels, Bel-

gium, pp. 810–815, 2012.

96. Yeom, S., I. Giacomelli, M. Fredrikson and S. Jha, "Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting", *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, Oxford, UK, pp. 268–282, 2018.

97. Squicciarini, A. C., A. Novelli, D. Lin, C. Caragea and H. Zhong, "From Tag to Protect: A Tag-Driven Policy Recommender System for Image Sharing", *2017 15th Annual Conference on Privacy, Security, and Trust (PST)*, Calgary, Alberta, Canada, pp. 337–33709, 2017.

98. Tran, L., D. Kong, H. Jin and J. Liu, "Privacy-CNH: A Framework to Detect Photo Privacy with Convolutional Neural Network Using Hierarchical Features", *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, Vol. 30, Phoenix Arizona, USA, 2016.

99. Squicciarini, A., C. Caragea and R. Balakavi, "Toward Automated Online Photo Privacy", *ACM Transactions on the Web (TWEB)*, Vol. 11, No. 1, p. 2, 2017.

100. Squicciarini, A. C., D. Lin, S. Sundareswaran and J. Wede, "Privacy Policy Inference of User-uploaded Images on Content Sharing Sites", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 27, No. 1, pp. 193–206, 2014.

101. Mosca, F. and J. Such, "ELVIRA: An Explainable Agent for Value and Utility-driven Multiuser Privacy", *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, Virtual Event, UK, p. 916–924, 2021.

102. Albertini, D. A., B. Carminati and E. Ferrari, "Privacy Settings Recommender for Online Social Network", *2016 IEEE 2nd International Conference on Collaboration and Internet Computing (CIC)*, Pittsburgh, PA, USA, pp. 514–521, 2016.

103. Carminati, B., E. Ferrari and A. Perego, "Rule-based Access Control for Social Networks", *On the Move to Meaningful Internet Systems*, Montpellier, France,

pp. 1734–1744, 2006.

104. Fong, P. W., "Relationship-Based Access Control: Protection Model and Policy Language", *Proceedings of the first ACM Conference on Data and Application Security and Privacy*, New York, NY, USA, pp. 191–202, 2011.

105. Yuan, L., J. Theytaz and T. Ebrahimi, "Context-Dependent Privacy-Aware Photo Sharing Based on Machine Learning", S. D. C. di Vimercati and F. Martinelli (Editors), *IFIP International Conference on ICT Systems Security and Privacy Protection*, Vol. 502, Rome, Italy, pp. 93–107, 2017.

106. Orekondy, T., B. Schiele and M. Fritz, "Towards a Visual Privacy Advisor: Understanding and Predicting Privacy Risks in Images", *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, pp. 3686–3695, 2017.

107. Li, Y., N. Vishwamitra, H. Hu and K. Caine, "Towards a Taxonomy of Content Sensitivity and Sharing Preferences for Photos", *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Honolulu, HI, USA, pp. 1–14, 2020.

108. Zhao, C., J. Mangat, S. Koujalgi, A. Squicciarini and C. Caragea, "PrivacyAlert: A Dataset for Image Privacy Prediction", *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16, Atlantic Georgia, USA, pp. 1352–1361, 2022.

109. Gilpin, L. H., D. Bau, B. Z. Yuan, A. Bajwa, M. Specter and L. Kagal, "Explaining Explanations: An Overview of Interpretability of Machine Learning", *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, Turin, Italy, pp. 80–89, 2018.

110. Riveiro, M. and S. Thill, "That's (not) the Output I Expected! On the Role of End User Expectations in Creating Explanations of AI Systems", *Artificial*

Intelligence, Vol. 298, p. 103507, 2021.

111. Langley, P., "Explainable, Normative, and Justified Agency", *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, Hawaii, USA, pp. 9775–9779, 2019.

112. Miller, T., "Explanation in Artificial Intelligence: Insights from the Social Sciences", *Artificial Intelligence*, Vol. 267, pp. 1–38, 2019.

113. Barredo Arrieta, A., N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado González, S. Garcia, S. Gil-Lopez, D. Molina, V. R. Benjamins, R. Chatila and F. Herrera, "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI", *Information Fusion*, Vol. 58, p. 82–115, 2019.

114. Selvaraju, R. R., M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, "Grad-cam: Visual Explanations from Deep Networks via Gradient-Based Localization", *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, pp. 618–626, 2017.

115. Xiao, T., Y. Xu, K. Yang, J. Zhang, Y. Peng and Z. Zhang, "The Application of Two-Level Attention Models in Deep Convolutional Neural Network for Fine-Grained Image Classification", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, Massachusetts, USA, 2015.

116. Shapley, L. S., "A Value for N-Person Games", *Classics in Game Theory*, Vol. 69, 1997.

117. Slack, D., S. Hilgard, E. Jia, S. Singh and H. Lakkaraju, "Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods", *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, New York, NY, USA, pp. 180–186, 2020.

118. Wohlin, C., P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell and A. Wesslén, *Experimentation in Software Engineering*, Springer Science & Business Media, 2012.

119. Likas, A., N. Vlassis and J. J. Verbeek, "The Global K-Means Clustering Algorithm", *Pattern Recognition*, Vol. 36, No. 2, pp. 451–461, 2003.

120. Johnson, S. C., "Hierarchical Clustering Schemes", *Psychometrika*, Vol. 32, No. 3, pp. 241–254, 1967.

121. Radford, A., J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision", *Proceedings of the 38th International Conference on Machine Learning*, Vol. 139, Virtual Event, pp. 8748–8763, 2021.

122. Miller, G. A., "WordNet: A Lexical Database for English", *Communications of the ACM*, Vol. 38, No. 11, pp. 39–41, 1995.

123. Schwartz, S. H., "An Overview of the Schwartz Theory of Basic Values", *Online readings in Psychology and Culture*, Vol. 2, No. 1, pp. 2307–0919, 2012.

124. Library, A. D., "Creative Commons Attribution 4.0 International License", `https://creativecommons.org/licenses/by/4.0`, accessed on Nov 25, 2013.

APPENDIX A: On the Figures of This Dissertation

Figures 1.1 and 1.2 are available under a Creative Commons Attribution 4.0 International license (CC BY 4.0) [124]. This dissertation respects the constraints and follows the license terms (i.e. gives appropriate credit and provides a link to the license) with respect to these figures. All the other figures presented in this dissertation are the author's original contributions, previously published in academic journals and conference proceedings. The figures that have emerged and have been transferred to the copyright publisher have been employed in compliance with the publisher's established "publication policy concerning the reuse of the author's own writings and graphics", as found on the publisher's official website.