Leveraging Machine Learning to Differentiate Legitimate versus Rogue Privacy Policies for Enhanced Decision Making

A. S. M. Kayes[1], Wenny Rahayu[1], Tharam Dillon[1], and Hooman Alavizadeh[1]

[1]Affiliation not available

April 16, 2024

Abstract

With the Internet of Things (IoT) generating vast amounts of data, privacy breaches have become increasingly prevalent, exposing individuals to serious risks such as identity theft and life-threatening situations. This research addresses the challenge of identifying cybersecurity threats and vulnerabilities leading to privacy breaches, as evidenced by recent cyber-attacks on Australian Medibank, Optus, and hospital networks. We propose a machine learning (ML)-based approach to distinguish between legitimate and rogue privacy policies, defining fundamental concepts of privacy, security, and access control in the context of personal, confidential, and sensitive information breaches. Our methodology introduces zero-privacy (ZP) and binary question-answer (QA) models to discern legitimate versus illegitimate actions or interests within privacy policies. Our experiments utilise natural language processing (NLP)-based ML models to analyse the linguistics of privacy policies. In experiments conducted on a dataset from the top 100 Forbes-listed companies, including 67 rogue policies, our privacy classification approach demonstrates reliability, accurately distinguishing between legitimate and rogue policies. With a dataset split of 90% for training and 10% for testing, our model achieves accuracy and precision exceeding 94% and 91%, respectively. Additionally, we evaluate the probability of ZP occurrences in organisations' privacy and service-level agreements, revealing significant privacy breach risks. Through case studies utilising our proposed binary QA model, we underscore the urgent need for enhanced privacy measures across various organisations' policies. Introducing a novel approach to access control, we specify permissions under conditions of legitimate and rogue privacy policies, exemplifying the applicability of our proposed access control mechanism through security policy modelling.

Leveraging Machine Learning to Differentiate Legitimate versus Rogue Privacy Policies for Enhanced Decision Making

A. S. M. Kayes, Wenny Rahayu, Tharam Dillon, and Hooman Alavizadeh

***Abstract*—With the Internet of Things (IoT) generating vast amounts of data, privacy breaches have become increasingly prevalent, exposing individuals to serious risks such as identity theft and life-threatening situations. This research addresses the challenge of identifying cybersecurity threats and vulnerabilities leading to privacy breaches, as evidenced by recent cyber-attacks on Australian Medibank, Optus, and hospital networks. We propose a machine learning (ML)-based approach to distinguish between legitimate and rogue privacy policies, defining fundamental concepts of privacy, security, and access control in the context of personal, confidential, and sensitive information breaches. Our methodology introduces zero-privacy (ZP) and binary question-answer (QA) models to discern legitimate versus illegitimate actions or interests within privacy policies. Our experiments utilise natural language processing (NLP)-based ML models to analyse the linguistics of privacy policies. In experiments conducted on a dataset from the top 100 Forbes-listed companies, including 67 rogue policies, our privacy classification approach demonstrates reliability, accurately distinguishing between legitimate and rogue policies. With a dataset split of 90% for training and 10% for testing, our model achieves accuracy and precision exceeding 94% and 91%, respectively. Additionally, we evaluate the probability of ZP occurrences in organisations' privacy and service-level agreements, revealing significant privacy breach risks. Through case studies utilising our proposed binary QA model, we underscore the urgent need for enhanced privacy measures across various organisations' policies. Introducing a novel approach to access control, we specify permissions under conditions of legitimate and rogue privacy policies, exemplifying the applicability of our proposed access control mechanism through security policy modelling.**

***Index Terms*—Privacy breaches, zero privacy, machine learning, natural language processing, legitimate policy, rogue policy.**

I. Introduction

THE proliferation of IoT leads to frequent and ongoing occurrences of cybersecurity incidents, resulting in an alarming surge in privacy breaches that jeopardise individuals. The misuse of individuals' personal, confidential, and sensitive information by cyber-criminals can lead to various fraudulent activities, ranging from credit card fraud to identity theft, causing both financial and emotional distress for those affected [1], [2]. The leakage of personal and sensitive information is a significant contributor to these detrimental effects, with potential consequences extending to life-threatening situations, especially when it involves patients' sensitive health information falling into unauthorised hands [3].

To address these concerns and protect individuals' privacy, adherence to regulations such as the Australian Privacy Act or the General Data Protection Regulation (GDPR) is essential [4]. An integral aspect of this effort is the examination of privacy policies implemented by organisations regarding the gathering, use, distribution, and storage of personal information. Regular reviews and updates of these policies are crucial to staying ahead of evolving cyber-threats and technological advancements.

In the context of internet platforms, which serve as crucial sources of information during critical situations like the COVID-19 pandemic, a concerning trend emerges [5]. A limited percentage of users take the time to read associated privacy statements, and cookie and service-level policies when connecting online. Research indicates that users' privacy awareness is low, leading to a lack of engagement with such policies. Classical access control and privacy solutions, including role-based access control (RBAC) [6] and context-aware access control (CAAC) [7] systems, face limitations in today's internet-driven environments due to the constraints of personal and sensitive information. The main risk stems from the insufficient adoption of these classical solutions, partly because they are predefined rule-based and associated with different contextual constraints [8].

To address these challenges, there is a need for innovative approaches to privacy and security control solutions that go beyond traditional rule-based systems. It is crucial to distinguish between legitimate and rogue privacy policies and integrate this classification concept into security policies. By doing so, organisations can limit access privileges based on the legitimacy of privacy policies, providing a more effective safeguard against potential misuse of personal information and enhancing overall privacy protection for individuals.

A. Research Motivation

Recent cyber incidents in Australia, such as personal and sensitive health information breaches in Medibank, Optus, and Victorian Hospital networks, underscore the need for organisations to improve their data privacy and security policies [5]. These breaches, resulting in unauthorised access to highly sensitive health information, contribute significantly to the

A. S. M. Kayes, Member, IEEE (a.kayes@latrobe.edu.au), Wenny Rahayu, Member, IEEE (w.rahayu@latrobe.edu.au), Tharam Dillon, Life Fellow, IEEE (t.dillon@latrobe.edu.au), and Hooman Alavizadeh, Member, IEEE (h.alavizadeh@latrobe.edu.au) are with the Department of Computer Science and Information Technology, La Trobe University, Melbourne, VIC 3086, Australia. (Corresponding author: A. S. M. Kayes).

Manuscript received April 01, 2024; revised April 30, 2024.

billions of dollars lost globally each year due to such incidents [9]. The scepticism towards the efficacy of access and privacy control systems is evident, as seen in the Australian My Health Record (MHR) scenario where millions of users opted out due to privacy and security concerns [3]. Inadequate access control undoubtedly plays a role in the surge of privacy breaches.

Organisations in various sectors collect and process vast amounts of personal information following their respective privacy policies. The complexity and volume of these policies pose a formidable challenge, potentially resulting in unintended leaks of personal information. Individuals, faced with extensive privacy policies filled with technical jargon, often overlook their thousands of lines of content before granting consent. Traditional privacy and access control approaches struggle to effectively manage these intricate policies.

To address these challenges, the adoption of an ML-based privacy assistant can help users and strengthen an organisation's overall data security, distinguishing legitimate versus rogue privacy policies. The overarching goal is to establish a fact-based understanding of how organisations specify their privacy policies and collect and process individuals' information for legitimate purposes. The contributions of this research are succinctly outlined as follows.

B. The Contributions

The primary objective of this research is to develop an ML-based approach to discern and classify legitimate versus rogue privacy policies. The overarching goal is to gain a factual understanding of how organisations articulate their privacy policies and handle individuals' information, ensuring legitimate interests are prioritised. The contributions of this research are succinctly outlined as follows:

- Introduction of a novel privacy approach accompanied by the fundamental conceptual underpinnings.
- Introduction of the binary QA and ZP models to categorise privacy policies into either legitimate or rogue.
- Empirical evaluation of the effectiveness of the proposed solution through a series of scientific experiments.
- Comprehensive exploration of the proposed privacy approach through illustrative case studies.
- Integration of the privacy classification concept into security policies to effectively control individuals' data.

C. Outline of the Article

The rest of the article is organised as follows. In Section 1, we articulate the motivation behind our research and highlight the novel contributions we bring to the field. Section 2 delves into the background of the research topic and provides an overview of related studies. Our novel approach to privacy classification is detailed in Section 3, outlining the methodology and concepts that distinguish our proposed model. Section 4 presents the results of various experiments and case studies conducted to showcase the effectiveness of our privacy classification approach. A user-centric approach to access control is introduced in Section 5. The paper concludes in Section 6, summarising key findings and proposing directions for future research.

II. Background and Related Research

In this section, we present comprehensive background information on access control mechanisms, alongside an exploration of relevant ML models and privacy approaches. To our knowledge, the existing landscape of access control has not yet integrated the concept of distinguishing rogue actions and policies from legitimate ones as part of security measures against privacy breaches.

A. Classical Access Control Background

Different access control mechanisms are already established as the treatment of privacy and security concerns in pervasive environments for protecting data and information resources from unauthorised users. The most crucial challenge is to specify the necessary policies to handle different types of information against privacy breaches. In the distributed cloud and IoT environments, while there are traditional RBAC standards available [6], an overlooked salient feature of CAAC is the specification of context-specific security and privacy policies [7], [8]. The diverse contextual information plays a vital role in specifying and enforcing CAAC policies. These classical RBAC and CAAC mechanisms have relied on the predefined policies that the security administrators specify statically. However, these mechanisms are not adequate in today's IoT-driven environments that are dynamic and distributed, which require a greater degree of autonomous decision-making.

We are currently living in the era of an interconnected world and the different types of contextual information which are associated with today's interconnected environments. For example, contextual information like the location from where the request came and the request time are taken into account to define the CAAC policies in different organisations [8]. Of the domain-specific contextual information, the different types of relationship context information are also associated with these policies, such as social or interpersonal relationships between users and data owners [10]. However, a significant number of individuals tend to overlook the privacy and service-level agreements employed by various organisations before granting access consent for the disclosure of personal, confidential, or sensitive information. Thus, privacy-specific relationship information also needs to be considered in the existing access control policies for better decision-making and to limit access privileges.

B. ML Models and Privacy Approaches

Liu et al. [11] surveyed different ML models and approaches to detect malicious IoT devices that are exposed to security risks. The comparative analysis between cryptographic approaches which are not suitable for many devices and systems is also discussed. They provided a comprehensive survey on ML technologies such as noncryptographic approaches for the detection of compromised IoT devices. The detection methods are classified into four categories such as abnormal device detection, device-specific pattern recognition, unsupervised device identification, and deep-learning-enabled device identification. These ML-based detection techniques are useful

to identify suitable ML models for distinguishing rogue actions and policies from legitimate ones.

A semantics-based privacy approach to IoT applications has been introduced [12], underlying the concepts of privacy by design (PbD) practices during the design phase of the system, such as privacy patterns, principles, guidelines, and strategies. The relevant privacy patterns have been considered across IoT systems and achieved through the development of an ontology for software developers. In accordance with the Australian Privacy Act and GDPR, these PbD measures can provide a useful tool to conceptualise the relevant legitimate versus malicious actions or interests within the privacy policies of various organisations.

Classification-based ML techniques and privacy models have been the focus of several studies [13]–[18]. Zimmeck and Bellovin [13] proposed an automatic classification solution aimed at enhancing privacy transparency online by analysing web privacy policies. This innovative approach reduces policy ambiguity, providing users with clearer insights into privacy practices on the web. Han and Shen [14] combined semi-supervised ML models with a graph-based approach to analyse phishing attacks, categorising unlabelled emails into different phishing campaign categories. Ravichander et al. [15] integrated computational and legal perspectives and proposed a QA model comprising thousands of questions about the privacy policies of mobile applications. Zaeem et al. [16] employed machine learning to digest privacy policies, resulting in an ML-based privacy analysis tool capable of summarising online privacy policies for web users. Recently, Saka et al. [17] employed unsupervised ML models to detect phishing scams by distinguishing malicious samples from benign ones. Alshamsan and Chaudhry [18] introduced a privacy framework that adheres to GDPR guidelines for data protection. Leveraging word-bags and scoring techniques, they implemented a web-based application to visually present risk-level reports for online privacy policies.

Identifying the constant privacy threats and emerging vulnerabilities that expose organisations to potential data breaches is an ongoing and challenging task. In recent years, the Australian Medibank, Optus, and hospital networks experienced security and privacy breaches perpetrated by cybercriminals, leading to the compromise of personal, confidential, and sensitive health information. Considering these challenges, existing ML models and privacy approaches are not adequate to distinguish legitimate versus rogue privacy policies. There is a need to discern legitimate and illegitimate actions within policies to decide whether a privacy policy is legitimate or rogue.

C. Privacy Control in the Internet of Things

Feng et al. [19] addressed the gap in designing effective privacy choices by constructing a comprehensive design space based on user-centric analysis. It offers a conceptual framework and taxonomy to guide practitioners in implementing legally compliant privacy choices, with a focus on privacy control in the IoT context.

Das et al. [20] outlined ongoing research on privacy assistants for IoT, aiming to empower users to regain control over their data. They discovered user-configurable settings for IoT resources (e.g., opt in/out, data erasure), assisting users in following their privacy expectations. They also discussed supporting personalised privacy settings through machine-learning-driven models of user preferences.

While Feng et al. [19] and Das et al. [20] offer valuable insights into designing privacy choices and personalised privacy assistants for IoT, they do not sufficiently address the distinction between legitimate and rogue privacy practices or provide measures to protect individuals' data against privacy breaches.

D. ML-Driven Access Control Approaches

Argento et al. [21] proposed an ML-based access control mechanism as the first line of defense, relying on users' behavioural patterns such as data volume and access frequency. Outchakoucht et al. [22] introduced a reinforcement-learning-based access control approach tailored for distributed IoT environments. Mayhew and Atighetchi [23] proposed a behaviour-based ML-driven access control system for anomaly detection, employing different ML algorithms to analyse HTTP requests and TCP connections.

Unlike traditional access control solutions, these ML-driven mechanisms offer automated decision-making capabilities. However, they currently lack the ability to effectively distinguish between rogue and legitimate actions, thus falling short in preventing privacy breaches.

E. Discussion

Existing access and privacy control solutions are not adequate to distinguish rogue actions from legitimate ones and to prevent privacy breaches. The static specification of access control policies, considering all privacy constraints, is a cumbersome and intricate administrative task. By leveraging ML, the distinction between rogue and legitimate actions within an organisation's privacy policies becomes achievable. Integrating this concept of distinguishing rogue versus legitimate actions into access control policies has the potential to limit access privileges, fortify protection against privacy breaches, and ultimately enhance decision-making capabilities. This advanced access and privacy control approach specifically aims to restrict permissions and prevent potential privacy breaches. A noteworthy aspect of this research involves the groundbreaking use of ML models to classify legitimate and rogue privacy policies. This innovative approach marks a significant stride towards more effective and dynamic privacy and access control mechanisms.

III. The Proposed Privacy Classification Approach

This section introduces the proposed privacy classification approach, including preliminary definitions and examples.

A. The Fundamental Concepts of Privacy, Security, and Access Control

It is essential to delineate between privacy [24], security [25], and access control [7], as they each represent distinct

facets of safeguarding personal, confidential, and sensitive information.

Definition 1: Privacy.

Privacy, as a fundamental principle, pertains to the individual's or entity's right to control who has access to their data, with strict enforcement of consent by the user.

Example 1: Privacy encapsulates the notion of consent, ensuring that data subjects retain control over their personal information. For instance, a social media platform may enable users to adjust their privacy settings, dictating who can view their posts or profile information. By strictly enforcing consent mechanisms, privacy measures aim to mitigate the risk of unauthorised access, such as misuse of data.

Privacy policies, expressed in natural language, can be utilised to enforce the rule in Definition 1.

Definition 2: Security.

Security relates to an unauthorised breach of access to information.

Example 2: Security encompasses a range of measures, such as the safeguarding of information from unauthorised access, alteration, or destruction, aimed at fortifying the integrity and confidentiality of data, for instance, the encryption of sensitive communications or the implementation of firewalls to thwart malicious cyber intrusions. Security breaches entail unauthorised access to personal, confidential, or sensitive information, potentially compromising its confidentiality or integrity.

Security protection through different security mechanisms such as encryption and intrusion detection can enforce the rule in Definition 2.

Definition 3: Access Control.

Access control relates to controlling who can access particular information under what conditions.

Example 3: Access control involves defining user permissions, authentication mechanisms, and enforcing policies to restrict unauthorised entry. For instance, an organisation might employ RBAC or CAAC [7] to assign privileges based on users' static roles or dynamic contexts. Access control policies serve as gatekeepers, ensuring that only authorised individuals can access confidential or sensitive information, thereby mitigating the risk of privacy breaches stemming from unauthorised access.

Access control policies expressed in natural language or using protocols can be used to enforce the rule in Definition 3.

By focusing on these three aspects, we aim to elucidate strategies for enhancing data protection and preserving individuals' rights to privacy. For instance, we explore the efficacy of privacy policies in governing data usage and the implementation of robust access control mechanisms to limit unauthorised access. In the following sections, we further discuss the privacy and access control issues surrounding personal, confidential, and sensitive information breaches.

B. Understanding Privacy Breaches

Privacy breaches can occur for a variety of reasons.

- Non-compliant privacy policies contain loopholes that allow the entity collecting information to intentionally use it beyond what the owner of the information intended.

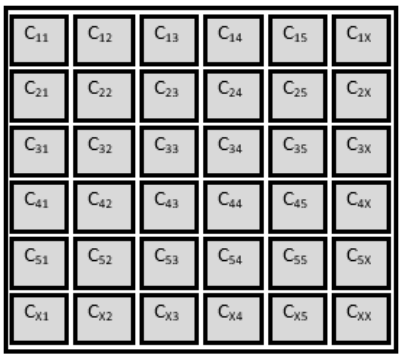

Fig. 1. An "*X by X*" surface to represent a legitimate policy with legitimate actions.

- Security breaches by external malicious actors involve stealing information and using it for their own purposes. In this situation, the fault lies with the original entity for failing to provide adequate security measures to prevent such breaches.
- Insider attacks by malicious individuals within the entity involve unauthorised access to information to misuse it in a way in which the original entity did not intend.

In this research, we focus on privacy breaches resulting from violations of privacy policies, differentiating between legitimate and illegitimate practices.

C. Legitimate Policy vs Rogue Policy

Definition 4: Legitimate Privacy Policy.

A legitimate privacy policy is represented as a 2-tuple relation, encompassing privacy rules (PR) and legitimate actions (LA).

$$LPP \ = \ \langle PR, \ LA \rangle \tag{1}$$

Example 4: A legitimate action or interest may involve collecting, storing, and/or disseminating personal, confidential, and/or sensitive information with authorised parties, for offering services to users.

Definition 5: Rogue Privacy Policy.

A rogue or illegitimate privacy policy is represented as a 2-tuple relation, combining privacy rules (PR) and rogue actions (RA).

$$RPP \ = \ \langle PR, \ RA \rangle \tag{2}$$

Example 5: A rogue or malicious or illegitimate action/interest may involve sharing or disseminating personal, confidential, and/or sensitive information with unauthorised parties, apart from offering services to users, possibly for marketing purposes.

These definitions and examples lay the groundwork for the subsequent development and application of the proposed privacy classification approach.

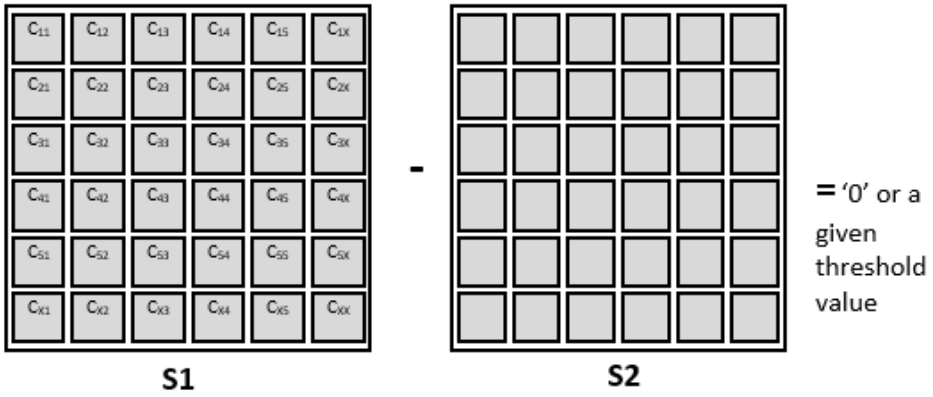

Fig. 2. Determining whether a privacy policy is legitimate or not.

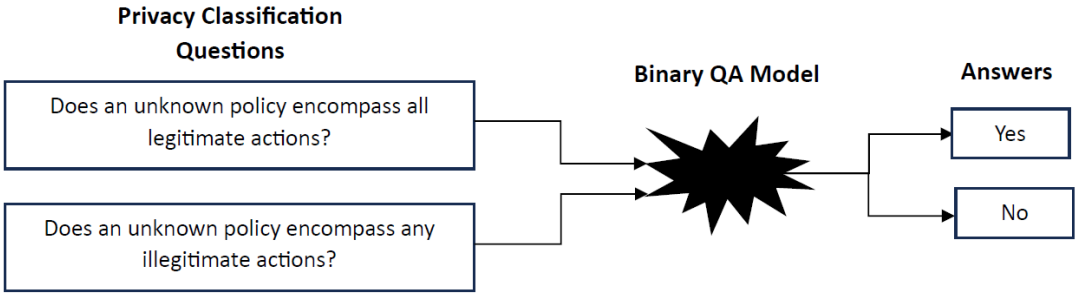

Fig. 3. An illustrative example of the privacy classification model.

D. Privacy Classification Approach

Figures 1 and 2 introduce a comprehensive framework for distinguishing between legitimate and rogue privacy policies.

Let us consider an 'X by X' surface. The linguistics of legitimate policies are distributed among all cells (i.e., from cells C_{11} to C_{XX} in Figure 1) on the surface. The cells represent the legitimate actions that were defined earlier. Consider two privacy policies: a known legitimate policy (Surface S1) and a new unknown policy (Surface S2) in Figure 2. The new privacy policy can be either legitimate or rogue. For a legitimate policy, the difference between the two surfaces ('S1 - S2') should be zero or fall within an acceptable threshold value. Otherwise, the new privacy policy can be considered rogue. The efficiency of decision-making (i.e., determining if a privacy policy is legitimate) can be calculated based on different 'X' values on the 'X by X' surface.

Definition 6: Zero-Privacy Model.

A zero-privacy (ZP) model is introduced based on the probability of ZP, denoted as P(ZP).

$$P(ZP) \; = \; count(C) \; on \; X \; by \; X \; surface \qquad (3)$$

The value of $P(ZP)$ will be equal to 100% when $count(C)$ = 1; otherwise, $P(ZP)$ will be less than 100%. An unknown or unlabelled privacy policy is considered legitimate when $count(C)$ = 1. The $count(C)$ value on an 'X by X' surface is 1 when all cells (C) on the surface encompass legitimate actions. In contrast, $count(C)$ = 0 when any one of the cells on the surface encompasses a rogue action.

Motivated by our binary causal question answering (QA) model [26], Figure 3 illustrates an example of our privacy classification model. It has two parts: a cause and an effect. For example, the privacy classification question "Does an unknown privacy policy encompass all legitimate actions?" has a cause (an unknown privacy policy) and an effect (all legitimate actions). The answer to the above binary causal question is either yes (1) or no (0). The proposed zero-privacy and binary QA models will assist in uncovering new and unknown/unlabelled privacy policies, categorising them as either $legitimate$ (1) or $rogue$ (0).

The inclusion of the privacy classification approach sets the stage for the experiment setup and provides a solid foundation for the subsequent experimental analysis.

IV. Experiments and Case Studies

In this section, we demonstrate the effectiveness of our proposed privacy classification approach through experiments and case studies.

A. Hypothesis and Use Cases

We consider the following hypothesis and two use cases to verify the hypothesis and train the ML models applied in our experiments.

Hypothesis. A larger 'X' value can produce better efficiency.

Use Case 1: We initiate the training of the ML models by using 50% labelled policies to fill as many cells as possible

on the surface. The remaining labelled policies can then be utilised for testing purposes. Efficiency will be higher if we increase the percentage of the training set relative to the testing set, essentially increasing the 'X' value.

Use Case 2: Initially, we employ a small labelled dataset (e.g., in our current experiments, considering 100 legitimate policies and 67 rogue policies). Gradually, we increase the number of policies in the labelled dataset. Efficiency will be higher as we increase the number of policies used to train the machine, essentially increasing the 'X' value.

B. Preparation of ML Models and Dataset

Our dataset consists of diverse privacy policies, comprising 67 illegitimate or rogue policies obtained from the online archive [27]. Additionally, we include 100 legitimate policies gathered from the websites of the top 100 organisations listed by Forbes magazine [27]. The complete dataset encompasses 167 privacy policies, ranging from Apple to Dark Blue Sea Group, totaling 287,016 words or 1,528,895 characters in length. Each policy has been manually assessed following our proposed privacy approach, adhering to the guidelines of the Australian Privacy Act and GDPR. We have labelled these policies '0' for rogue and '1' for legitimate.

We leverage five robust NLP-based ML techniques - BERT, Distil BERT, RoBERTa Tokenizer, Albert Tokenizer, and French Language models - to discern and distinguish rogue privacy policies from legitimate ones. The selection of these transformer-based text classification models is underpinned by their proven efficacy in handling complex text data and language understanding tasks. BERT and Distil BERT, known for their deep contextual embeddings, provide comprehensive semantic understanding. RoBERTa Tokenizer excels in fine-grained language modeling, while Albert Tokenizer's lightweight yet efficient architecture complements the ensemble. The incorporation of French Language models ensures domain-specific relevance and accuracy. By leveraging the collective power of these models, we enhance the capability of our proposed privacy approach to analyse intricate policy documents/statements, improve decision-making, and ultimately fortify privacy measures.

C. Experiment Setup

We conducted our experiments in the Python 3 Google Colab environment, utilising resources comprising 12.72 GB of RAM and 68.40 GB of disk space.

We implemented the following procedure to classify privacy policies into legitimate and rogue, using transformer models such as BERT, Distill BERT, and RoBERTa.

1) **Installation of Necessary Libraries:**
   - We began by installing essential Python libraries, including torch and pandas.
2) **Dataset Features:** Our dataset consists of two columns:
   - Policy: Textual content representing privacy policies from various organisations.
   - Label: A binary value indicating whether the policy is legitimate (1) or rogue (0).
3) **Loading and Preprocessing Dataset:**
   - We loaded the dataset from a CSV file into the pandas DataFrame, performed preprocessing tasks, and converted it into a format compatible with the Hugging Face libraries.
   - To address class imbalance issues, we balanced the dataset and converted it into the appropriate format for the Hugging Face libraries.
   - The dataset was split into training and testing sets for model evaluation purposes.
   - We utilised the transformer tokenizer to preprocess the textual content of privacy policies, ensuring that sequences did not exceed the maximum input length of the transformer models.
   - We batched the preprocessed data, enabling the parallel processing of multiple elements.
4) **Model Training and Evaluation:**
   - We defined the base model architecture, optimizer, learning rate scheduler, and device for training.
   - In cases where certain weights were not initialised from the model checkpoints, we fine-tuned the models on downstream tasks to facilitate their use for predictions and inferences.
   - Evaluation metrics (accuracy, precision, recall, and F1 score) were calculated after each training epoch, and the best-performing model was tracked.
   - The state dictionary of the best model was saved to a file, and an inference was made using the fine-tuned model.
   - The metric results were recorded and analysed post-training along with the training loss and validation accuracy.

This systematic approach allowed us to effectively classify whether a privacy policy is legitimate or not using transformer models, providing insights into the performance and capabilities of various architectures.

D. Experiment Results

Figure 4 presents the results of the experiments where the training dataset comprises 50% of the dataset, and the testing dataset comprises the remaining 50%. The results indicate that the BERT and RoBERTa Tokenizer models have an identical F1 score, accuracy, and precision values, surpassing the performance of the other three ML techniques in distinguishing legitimate versus rogue privacy policies. Specifically, the BERT model exceeds 82% accuracy and 88% precision. In summary, various ML models in this setup produce good results. For instance, our proposed approach demonstrates the ability to distinguish rogue policies from legitimate ones, with an F1 score exceeding 0.84 using both BERT and RoBERTa Tokenizer models.

The dataset undergoes multiple passes through these five ML models to optimise learning, adjusting the number of epochs. Another set of experiment results with a 90% training and 10% testing split is summarised in Table I. It illustrates the accuracy and precision of the different ML models on the dataset, with the Distil BERT model achieving exceptional

BERT (Accuracy: 0.8182, Precision: 0.8776, Recall: 0.8113, F1 Score: 0.8431)

Epoch	Training Loss	Validation Accuracy
1	0.2899	0.8333
2	0.2773	0.8333
3	0.1856	0.8333
4	0.2172	0.8889
5	0.129	0.8889
6	0.1236	0.8889

Distil BERT (Accuracy: 0.7045, Precision: 0.6800, Recall: 0.9623, F1 Score: 0.7969)

Epoch	Training Loss	Validation Accuracy
1	0.0222	0.8333
2	0.0156	0.8333
3	0.0118	0.7222
4	0.0085	0.7222
5	0.0075	0.6667
6	0.0055	0.8333

Roberta Tokenizer (Accuracy: 0.8182, Precision: 0.8776, Recall: 0.8113, F1 Score: 0.8431)

Epoch	Training Loss	Validation Accuracy
1	0.2899	0.8333
2	0.2773	0.8333
3	0.1856	0.8333
4	0.2172	0.8889
5	0.129	0.8889
6	0.1236	0.8889

Albert Tokenizer (Accuracy: 0.7159, Precision: 0.6842, Recall: 0.9811, F1 Score: 0.8062)

Epoch	Training Loss	Validation Accuracy
1	0.0125	0.6111
2	0.0309	0.6667
3	0.0091	0.6667
4	0.0065	0.6667
5	0.0068	0.6667
6	0.0062	0.7222

French Language (Accuracy: 0.6023, Precision: 0.6023, Recall: 1.0000, F1 Score: 0.7518)

Epoch	Training Loss	Validation Accuracy
1	0.6023	0.6667
2	0.5831	0.7222
3	0.5318	0.7222
4	0.5029	0.7222
5	0.4858	0.7778
6	0.4565	0.7778

Fig. 4. Training loss and validation accuracy of various NLP-based ML models (training dataset: 50%, testing dataset: 50%).

TABLE I
A SUMMARY OF ACCURACY AND PRECISION VALUES WITH DIFFERENT ML MODELS (TRAINING DATASET: 90%, TESTING DATASET: 10%).

ML Techniques	Accuracy	Precision
BERT	0.8889	0.8462
Distil BERT	0.9444	0.9167
RoBERTa Tokenizer	0.8889	0.8462
Albert Tokenizer	0.7778	0.8889
French Language	0.6111	0.6111

accuracy and precision values. The model achieves an accuracy exceeding 94%, showcasing its effectiveness in accurately classifying privacy policies. Furthermore, the precision of the model surpasses 91%, highlighting its ability to precisely identify and differentiate between legitimate and rogue policies. These results affirm the robustness and reliability of our proposed ML-based privacy classification approach.

As illustrated in Figure 5 and detailed in Table I, our proposed privacy classification approach exhibits commendable proficiency in identifying rogue policies. This efficacy is particularly noteworthy given a dataset split of 10% for testing and 90% for training. The inclusion of data labels in Figure 5 enhances clarity, providing a visual representation of the accuracy across all five ML models. It is noteworthy that the accuracy and precision of these ML models greatly improved as they were well-trained on a substantial number of privacy policies, specifically with a training dataset comprising a higher split compared to the testing dataset.

Overall, through different experiments, we evaluated the proposed privacy approach on a dataset containing 100 legitimate and 67 rogue privacy policies. In the following section, using different case studies, we evaluate legitimate versus malicious actions or interests. In the earlier sections, a legitimate privacy policy is represented as a 2-tuple relation, including privacy rules and legitimate actions/interests (e.g., collecting, storing, or disseminating data where personal and sensitive information is involved with authorised parties). A rogue privacy policy is represented as a 2-tuple relation, including privacy rules and rogue actions/interests (e.g., sharing personal and sensitive information with unauthorised parties, apart from offering services to people for marketing purposes).

E. Walkthrough and Case Studies

In this section, we detail our proposed zero-privacy and binary QA models to differentiate various privacy policies, categorising them as either $legitimate$ (1) or $rogue$ (0).

Case Study 1: The Apple Privacy Policy. Let us examine Apple's privacy policy. A snapshot of the policy states: *Your privacy is important to Apple. So we've developed a Privacy Policy that covers how we collect, use, disclose, transfer, and store your information. What personal information we collect - we use personal information to help us create, develop, operate, deliver, and improve our products, services, content, and advertising. We may use your personal information, including your date of birth, to verify identity, assist with the identification of users, and to determine appropriate services.*

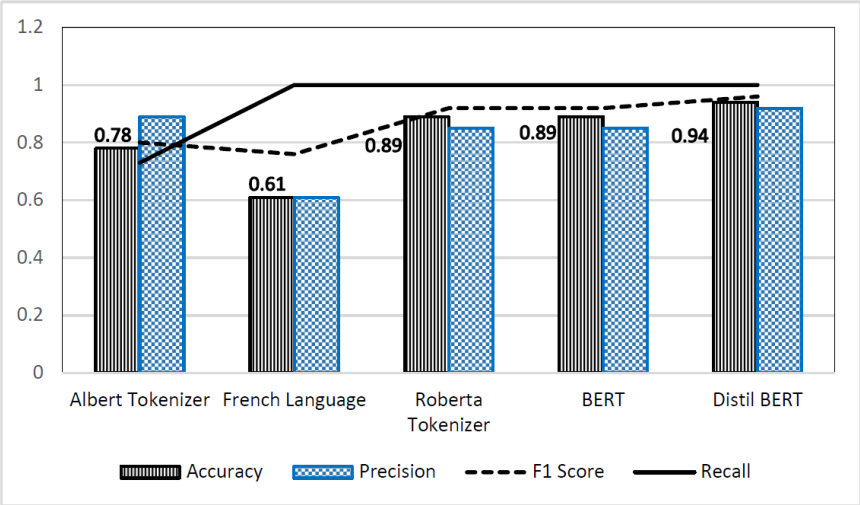

Fig. 5. Experiment results with different ML models (training dataset: 90%, testing dataset: 10%).

For example, we may use date of birth to determine the age of Apple ID account holders.

Apple's privacy policy clearly articulates the collection of personal information for purposes such as user verification, service provision, product enhancement, and advertising. Importantly, it specifies that sensitive information like gender orientation or political beliefs is not gathered.

Utilising our zero-privacy (ZP) model, we calculate a $count(C)$ value of 1 for Apple's privacy policy, resulting in a $P(ZP)$ of 100%. The binary QA model, as illustrated in Figure 3, affirms that all privacy classification questions indicate the absence of illegitimate actions/interests in Apple's privacy policy. Therefore, based on our privacy approach, Apple's privacy policy is classified as $legitimate$.

Case Study 2: The Dark Blue Sea Privacy Policy. Let us consider a snapshot of Dark Blue's privacy policy: *The Dark Blue Sea group of companies, part of the Photon Group, offers an exciting range of online products and services. This Privacy Policy has been drafted to comply with the national privacy principles set out in the Australian Privacy Act 1988 (Cth). We use personal information that we collect to conduct our business of delivering online products. In order to do this, we share your personal information with other companies.*

Regarding the Dark Blue Sea privacy policy, it involves the use of clients' personal information for business operations and product delivery. According to their policy, they share this personal information with other companies without obtaining explicit consent from clients.

Applying our proposed 'ZP' model, we calculate $count(C)$ = 0 for the Dark Blue Sea's privacy policy, resulting in a $P(ZP)$ of less than 100%. Using the binary QA model illustrated in Figure 3, all privacy classification questions yield negative responses, indicating that this policy involves illegitimate actions and violates their clients' privacy. For example, the statement "*we share your personal information with other companies*" is considered an illegitimate interest/action, as it does not explicitly outline the purpose of sharing confidential and sensitive personal information.

F. Limitations and Future Experiments

This section outlines the limitations of our current experiments and how we can extend the setup for future experiments.

1) Limitations:

- **Limited Dataset and Manual Labelling Process**: The dataset is currently comprised only of labelled policies, and the labelling process was carried out based on our proposed privacy classification approach, aligning with the Australian Privacy Act and GDPR. The labelling of policies as '1' ($legitimate$) or '0' ($rogue$) was done manually, following the definitions and approach established in Section 3.
- **Limited ML Models**: The experiments are currently restricted to five NLP-based text classification techniques. To enhance the robustness of the proposed privacy classification approach, we plan to extend the experiment setup by incorporating large language models [28]. This expansion will enable us to audit the collection and distribution of sensitive information outlined in privacy policies, thus contributing significantly to the depth and effectiveness of our research.

2) Future Experiments:

- **Categorisation of Privacy Policies Beyond Binary Classifications**: We intend to refine the granularity of our privacy policy classification by categorising them at various levels of sensitivity. This includes distinguishing between less sensitive personal data, personally identifiable information, credit card numbers, critical business data, and more sensitive health information. As part of this approach, we aim to label privacy policies across multiple categories, moving beyond the simplistic binary '0' and '1' classifications. This expanded classification will incorporate considerations of the impact of breaches on different types of sensitive information, thereby providing a more comprehensive framework for understanding and addressing privacy concerns.

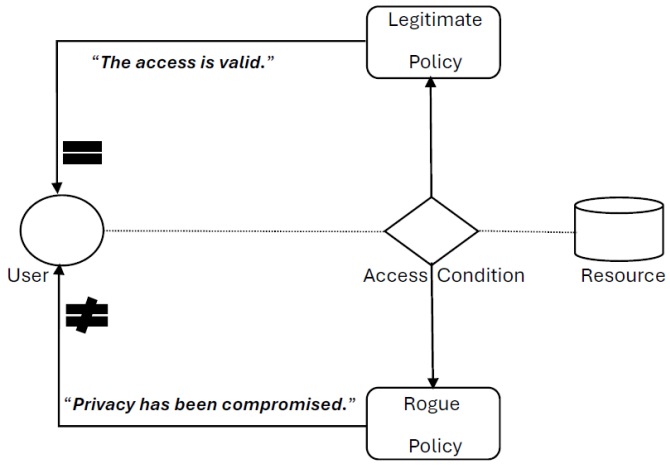

Fig. 6. A user-centric approach to access control.

- **Automation of Labelling Process**: To streamline the labelling process, we will develop a Python script to automate the categorisation of organisations' policies
- **Model Improvement and Handling Unknown Policies**: We aim to enhance the precision and accuracy of our ML models by incorporating a larger and more diverse dataset. Additionally, we plan to extend the proposed approach to accommodate decisions regarding both labelled/known and unlabelled/unknown privacy policies.

G. Discussion

In summary, our experimental endeavours have yielded significant insights into the landscape of privacy policies and practices within various organisations.

By effectively discerning between legitimate and rogue privacy policies, we have provided users with a valuable tool to navigate the complexities of organisations' privacy and service-level agreements. Furthermore, our analysis has enabled us to quantify the likelihood of privacy breaches, denoted as $P(ZP)$, shedding light on the prevalent challenges facing individuals' personal data security. Drawing from compelling case studies, we have illustrated the widespread non-adherence to privacy standards by organisations, emphasising the urgent need for improved safeguards. Moving forward, we will address the limitations of our study and explore avenues for future research and extension.

In light of these findings, we advocate a user-centric approach to access control in the following section, placing individuals at the forefront of privacy protection efforts.

V. User-Centric Access Control Approach

The extensive collection of personal, confidential, and sensitive information by organisations necessitates a proactive approach to security policies (i.e., access control policies). Providing the ability to increase access control granularity and decision-making precision becomes imperative for enhancing data security. By advocating a user-centric approach to access control (see Figure 6), we can empower individuals to control their data and foster a safer, more privacy-respecting online environment.

We incorporate the classification of legitimate versus rogue privacy concepts into security policies and propose a user-centric approach to access control.

The proposed access control can prevent users' access to resources (e.g., data and services). This can effectively limit access privileges and hinder decision-making capability. Access is granted when the access condition satisfies the legitimate actions outlined in the privacy policy and denied when the privacy policy is violated.

A. Modelling Access Control Policies

We model security policies to control unauthorised access to individuals' data. We enhance our earlier context and policy models [7], [8] by incorporating legitimate and illegitimate actions as contextual conditions. Specifically, we integrate the privacy classification concept into the security policies as a crucial element of the access control system.

TABLE II
An example access control policy for Dark Blue Sea resources and services to prevent privacy breaches.

If User() is 'dark-blue-clients' $\wedge$ Access Condition() encounters an illegitimate action or interest in the privacy statement, resulting in the identification of the privacy as "Rogue Policy" $\wedge$ Resource() is 'online-navigation' $\rightarrow$ Access permission is 'denied' with the reason "Privacy has been compromised."

For instance, the security policy in Table II demonstrates how access control can deny users' requests when the personal information of individuals has been compromised. This user-specific access control policy can be applied to manage clients' personal information for business operations and product delivery, as demonstrated in the Dark Blue Sea Privacy Policy in Case Study 2.

The privacy classification concept has been incorporated into the aforementioned security policy. Through this transparent and granular control mechanism, users can exert greater agency over their data.

VI. Conclusion and Promising Future Research Directions

In this paper, we introduced a novel privacy approach aimed at classifying legitimate versus rogue privacy policies, employing binary QA and ZP models as our methodology. Following our approach, we employed five transformer-based text classification techniques to discern and distinguish rogue policies from legitimate ones. We utilised a publicly available dataset comprising 167 privacy documents and passed them through these ML models multiple times to optimise learning. With a dataset split of 90% for training and 10% for testing, the Distil BERT model outperformed the other four models in terms of precision, accuracy, and F1 score, demonstrating the reliability of our proposed privacy classification approach.

Finally, we introduced a user-centric approach to access control, which empowers users to regulate access to their personal, confidential, and sensitive information. Through the specification of a security policy, our approach demonstrates greater transparency and control over data access, effectively mitigating the risks associated with privacy breaches. By placing individuals at the forefront of privacy protection efforts, this approach ensures greater transparency and control over data access, thereby mitigating the risks associated with privacy breaches within IoT-driven digital environments.

- **Privacy Awareness**: The prevailing tendency among individuals to overlook the content of privacy statements/policies before granting consent poses a significant risk of privacy breaches. It is crucial to address this issue by promoting education and actively raising awareness among individuals about the importance of safeguarding their information. We aim to mitigate the potential risks associated with the leakage of personal and sensitive information. Through these efforts, we strive to alleviate the adverse consequences of privacy breaches and protect individuals who may be vulnerable to misuse.
- **Developing an ML-Based Privacy Assistant for Threat Identification**: Continuously identifying privacy threats and emerging vulnerabilities is an ongoing research imperative for safeguarding organisations against potential breaches. Our upcoming initiative involves the development of a software prototype grounded in diverse ML techniques. The goal is to automate privacy classification and provide robust protection against privacy breaches. We aim to calculate users' confidence scores and non-binary classifications of privacy and service-level policies, considering the varying sensitivities of information.
- **Enhancing Privacy Protection**: It is imperative to not only consider the probability of a privacy breach occurring but also to delve deeper into the impact of such breaches. This entails assessing the significance of the information that was the subject of the breach, such as credit card numbers, health records, or other sensitive data. Understanding the potential consequences of data exposure is crucial for devising more effective mitigation strategies and enhancing overall privacy protection measures. By incorporating the impact of breaches on different types of sensitive information, future research can provide a more comprehensive understanding of the risks posed by privacy breaches and inform the development of targeted access and risk management approaches. Additionally, exploring methodologies to quantify the severity of privacy breaches based on the nature and sensitivity of the compromised data can contribute to the refinement of risk assessment frameworks and the implementation of more tailored access control policies and other security measures.

REFERENCES

[1] P. Mayer, Y. Zou, F. Schaub, and A. J. Aviv, "Now I'm a bit angry: Individuals' awareness, perception, and responses to data breaches that affected them," in *USENIX Security*, 2021, pp. 393–410.

[2] S. Kemp, "Exploring public cybercrime prevention campaigns and victimization of businesses: A Bayesian model averaging approach," *Computers & Security*, vol. 127, p. 103089, 2023.

[3] Z. Hollo and D. E. Martin, "An equitable approach to enhancing the privacy of consumer information on My Health Record in Australia," *Health Information Management Journal*, vol. 52, no. 1, pp. 37–40, 2023.

[4] L. Kyi, S. Ammanaghatta Shivakumar, C. T. Santos, F. Roesner, F. Zufall, and A. J. Biega, "Investigating deceptive design in GDPR's legitimate interest," in *ACM CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–16.

[5] K. Michael and R. Abbas, "What happens to COVID-19 data after the pandemic? Socio-technical lessons," *IEEE Transactions on Technology and Society*, vol. 3, no. 4, pp. 242–247, 2022.

[6] D. F. Ferraiolo, R. Sandhu, S. Gavrila, D. R. Kuhn, and R. Chandramouli, "Proposed NIST standard for role-based access control," *ACM Transactions on Information and System Security*, vol. 4, no. 3, pp. 224–274, 2001.

[7] A. Kayes, J. Han, W. Rahayu, T. Dillon, M. S. Islam, and A. Colman, "A policy model and framework for context-aware access control to information resources," *The Computer Journal*, vol. 62, no. 5, pp. 670–705, 2019.

[8] A. Kayes, W. Rahayu, P. Watters, M. Alazab, T. Dillon, and E. Chang, "Achieving security scalability and flexibility using fog-based context-aware access control," *Future Generation Computer Systems*, vol. 107, pp. 307–323, 2020.

[9] H. Tao, M. Z. A. Bhuiyan, M. A. Rahman, G. Wang, T. Wang, M. M. Ahmed, and J. Li, "Economic perspective analysis of protecting big data security and privacy," *Future Generation Computer Systems*, vol. 98, pp. 660–671, 2019.

[10] A. Kayes, J. Han, and A. Colman, "An ontological framework for situation-aware access control of software services," *Information Systems*, vol. 53, pp. 253–277, 2015.

[11] Y. Liu, J. Wang, J. Li, S. Niu, and H. Song, "Machine learning for the detection and identification of Internet of Things devices: A survey," *IEEE Internet of Things Journal*, vol. 9, no. 1, pp. 298–320, 2021.

[12] L. Alkhariji, S. De, O. Rana, and C. Perera, "Semantics-based privacy by design for Internet of Things applications," *Future Generation Computer Systems*, vol. 138, pp. 280–295, 2023.

[13] S. Zimmeck and S. M. Bellovin, "Privee: An architecture for automatically analyzing web privacy policies," in *USENIX Security*, 2014, pp. 1–16.

[14] Y. Han and Y. Shen, "Accurate spear phishing campaign attribution and early detection," in *31st Annual ACM Symposium on Applied Computing*, 2016, pp. 2079–2086.

[15] A. Ravichander, A. Black, S. Wilson, T. Norton, and N. Sadeh, "Question answering for privacy policies: Combining computational and legal perspectives," in *2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing*, 2019, pp. 4947–4958.

[16] R. N. Zaeem, S. Anya, A. Issa, J. Nimergood, I. Rogers, V. Shah, A. Srivastava, and K. S. Barber, "Privacycheck's machine learning to digest privacy policies: Competitor analysis and usage patterns," in *IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*. IEEE, 2020, pp. 291–298.

[17] T. Saka, K. Vaniea, and N. Kökciyan, "Context-based clustering to mitigate phishing attacks," in *ACM Workshop on Artificial Intelligence and Security*, 2022, pp. 115–126.

[18] A. R. Alshamsan and S. A. Chaudhry, "A GDPR compliant approach to assign risk levels to privacy policies," *Computers, Materials & Continua*, vol. 74, no. 3, 2023.

[19] Y. Feng, Y. Yao, and N. Sadeh, "A design space for privacy choices: Towards meaningful privacy control in the Internet of Things," in *ACM CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–16.

[20] A. Das, M. Degeling, D. Smullen, and N. Sadeh, "Personalized privacy assistants for the Internet of Things: Providing users with notice and choice," *IEEE Pervasive Computing*, vol. 17, no. 3, pp. 35–46, 2018.

[21] L. Argento, A. Margheri, F. Paci, V. Sassone, and N. Zannone, "Towards adaptive access control," in *32nd Annual IFIP Conference on Data and Applications Security and Privacy*, F. Kerschbaum and S. Paraboschi, Eds. Springer, 2018, pp. 99–109.

[22] A. Outchakoucht, E. Hamza, and J. P. Leroy, "Dynamic access control policy based on blockchain and machine learning for the Internet of Things," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 7, pp. 417–424, 2017.

[23] A. Adler, M. J. Mayhew, J. Cleveland, M. Atighetchi, and R. Greenstadt, "Using machine learning for behavior-based access control: Scalable anomaly detection on TCP connections and HTTP requests," in *IEEE Military Communications Conference*. IEEE, 2013, pp. 1880–1887.
[24] M. Hecker, T. S. Dillon, and E. Chang, "Privacy ontology support for e-commerce," *IEEE Internet Computing*, vol. 12, no. 2, pp. 54–61, 2008.
[25] R. Tourani, S. Misra, T. Mick, and G. Panwar, "Security, privacy, and access control in information-centric networking: A survey," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 1, pp. 566–600, 2017.
[26] H. Kayesh, M. S. Islam, J. Wang, S. Anirban, A. Kayes, and P. Watters, "Answering binary causal questions: A transfer learning based approach," in *2020 International Joint Conference on Neural Networks*, 2020, pp. 1–9.
[27] K. Rekanar and M. Boldt, "Privacy policies from legitimate and rogue web sites," in *DOI: https://doi.org/10.5281/zenodo.242385*. Dataset updated on May 21, 2018.
[28] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson *et al.*, "Extracting training data from large language models," in *USENIX Security*, 2021, pp. 2633–2650.

VII. Biographies

Dr. A. S. M. Kayes is a Senior Lecturer in Cybersecurity at La Trobe University, Australia. His research interests encompass various areas within cybersecurity, such as data security, access control, fog and cloud security, cyber incidents, and data/privacy breaches. Over the past decade, he has made significant contributions to the field, with more than 75 research articles published in international journals and conference proceedings. He has an h-index of 25 and over 2,400 citations (Google Scholar). He is a member of the Australian Computer Society and IEEE.

Prof. Wenny Rahayu is a Professor and Dean of the School of Computing, Engineering, and Mathematical Sciences at La Trobe University, Australia. Her research interests include data privacy, big data integration and management, and access control. Over the past 15 years, she has published 2 books and more than 260 research articles in international journals and conference proceedings. She has an h-index of 40 and over 8,000 citations (Google Scholar). She is a member of the Australian Computer Society and IEEE.

Prof. Tharam Dillon is an adjunct Professor at La Trobe University, Australia. He has published 8 authored books and more than 500 research papers in international journals and conference proceedings. His research works have been widely cited and therefore have considerable impact. He has an h-index of 67 and over 20,500 citations (Google Scholar), which puts him in the top percentile of researchers globally. He is currently a Life Fellow of the IEEE and a Fellow of the Australian Computer Society and the Institution of Engineers Australia.

Dr. Hooman Alavizadeh is currently a Lecturer in Cybersecurity in the Department of Computer Science and Information Technology at La Trobe University, Australia. His research interests include cloud and network security, security modelling and analysis, moving target defence, cryptography, and cyber situation-awareness. He has an h-index of 12 and over 600 citations (Google Scholar). He is a member of the Australian Computer Society and IEEE.