YaleNUSCollege

GHOSTS IN THE MACHINE: APPLYING & EVALUATING EXPLAINABLE AI TECHNIQUES IN LEGAL DECISION MAKING

TRISTAN KOH LY WEY

Capstone Final Report for Yale-NUS BSc (Honours) in Mathematical, Computational and Statistical Sciences & LLB (Honours) Double Degree Program

Supervised by:

Professor Michael Choi and Professor Simon Chesterman

AY 2022/2023

Yale-NUS College Capstone Project

DECLARATION & CONSENT

1. I declare that the product of this Project, the Thesis, is the end result of my own work and that due acknowledgement has been given in the bibliography and references to ALL sources be they printed, electronic, or personal, in accordance with the academic regulations of Yale-NUS College.

2. I acknowledge that the Thesis is subject to the policies relating to Yale-NUS College Intellectual Property (Yale-NUS HR 039).

ACCESS LEVEL

3. I agree, in consultation with my supervisor(s), that the Thesis be given the access level specified below: [check one only]

   o Unrestricted access
   Make the Thesis immediately available for worldwide access.

   o Access restricted to Yale-NUS College for a limited period
   Make the Thesis immediately available for Yale-NUS College access only from _____ (mm/yyyy) to _____ (mm/yyyy), up to a maximum of 2 years for the following reason(s): (please specify; attach a separate sheet if necessary):
   _____.

   After this period, the Thesis will be made available for worldwide access.

   o Other restrictions: (please specify if any part of your thesis should be restricted)
   _____
   _____

_____
Name & Residential College of Student

_____ _____
Signature of Student Date

_____ _____
Name & Signature of Supervisor Date

Acknowledgements

This capstone would not be possible if not for the gracious provision of the following individuals and groups, and most of all, God.

他对我说:"我的恩典够你用的,因为我的能力是在人的软弱上显得完全."所以,我更喜欢夸耀自己的软弱,好使基督的能力覆庇我. - 哥林多后书12:9

YALE-NUS COLLEGE

Abstract

B.Sc (Hons) and L.L.B (Hons)

Ghosts in the Machine: Applying & Evaluating Explainable AI in Legal Decision Making

by Tristan KOH

This capstone assesses the explainability of the visualisations of machine learning models that were trained to identify data practices within apps' data privacy policies.

Natural language processing (NLP) techniques are now much more advanced and can potentially automate some areas of substantive legal reasoning and writing, but has come at the cost of increased opacity of the models. Designing Explainable AI (XAI) is a method to combat this opacity, but it is unclear whether these explanations are effective across the different users and use cases of the models. The stakeholders in data privacy regulation can also benefit from the use of XAI. There is a gap in research of XAI assessments generally, and XAI's particular application in data privacy.

This capstone uses this data privacy context to train and evaluate AI classifiers that are able to detect data practices within apps' data privacy

practices. XAI techniques are applied on these models to produce visualisations that aim to explain these predictions to these stakeholders that may not be experts in data science. Finally, the effectiveness of these visualisations are assessed by surveying the general public along differing metrics of explainability.

"Any sufficiently advanced technology is indistinguishable from magic."

Arthur C. Clarke

Contents

List of Tables

List of Figures

1 Introduction

1.1 Motivation and significance

A question was posed in 2017: "Will there be a software service for the generation of explanations and arguments in law that will assist in structuring explanations of answers and supportive legal arguments?" (Ashley, 2017). This question seems close to getting an answer in 2023 with the release of ChatGPT (OpenAI, 2022) to the general public. Consider the following response from ChatGPT:

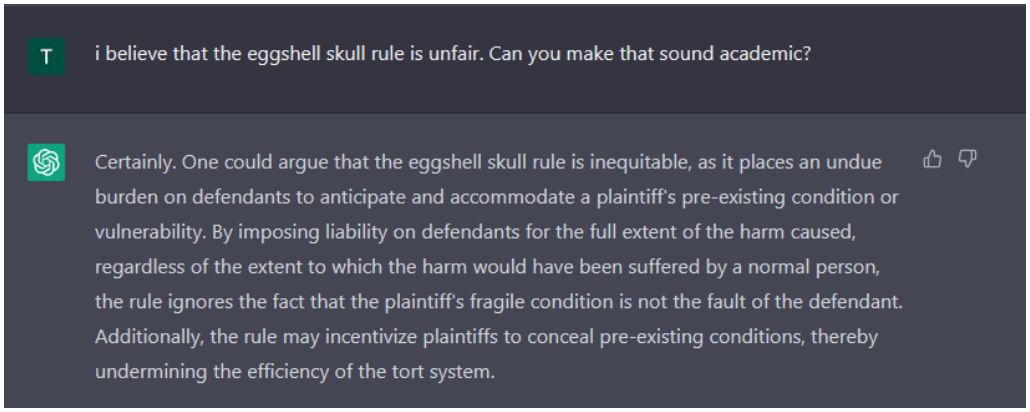


FIGURE 1.1: Conversation with ChatGPT about the eggshell skull rule.

Indeed, GPT-4, successor of ChatGPT and released a mere 4 months after ChatGPT, scored on the 90th percentile on the Uniform Bar Exam which is significantly in excess of the passing mark to practice law in the US (Katz et al., 2023a).

ChatGPT and GPT-4 are the most recent AI models that uses advanced natural language processing (NLP). NLP is a branch of AI that gives computers the ability to understand text and spoken words as how human beings would understand (IBM, 2022). While NLP techniques have substantially increased in performance in recent years, it has come at the cost of the explainability of their predictions because of models that are architecturally more complex (El Zini and Awad, 2022). This issue of explainability is exemplified by the last question I posed to ChatGPT about its view on the egg-shell skull rule:

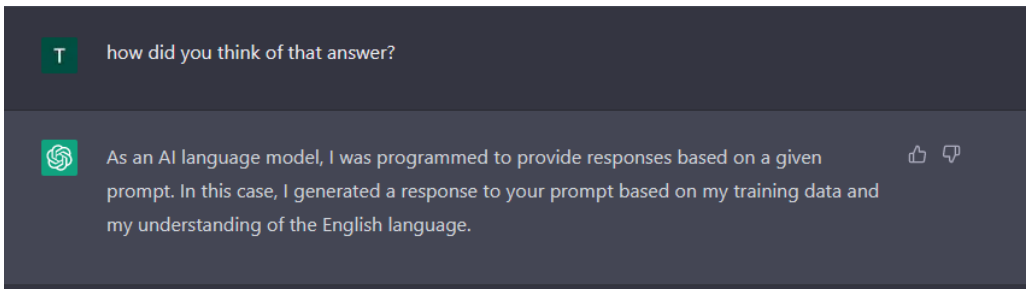


FIGURE 1.2: Conversation with ChatGPT about the egg-shell skull rule.

ChatGPT does not seem to be able to explain its views like how a typical human would[1]. This lack of explainability could be a significant hindrance towards NLP's greater adoption within the legal industry because the lawyer that uses these models ultimately bear the legal responsibility of ensuring that the analysis is legally sound. In Singapore, a legal practitioner must act with reasonable diligence and competence in the provision of services to the client[2]. A lawyer that relies on the analysis of

[1]Further prompting led ChatGPT to provide a list of academic papers that support its view. But there are issues of hallucination, where ChatGPT generates answers about facts that are false or non-existent (Alkaissi and McFarlane, 2023).

[2]According to r5(2)(c) of the Legal Profession (Professional Conduct) Rules 2015.

legal tech tools but does not understand how the analysis was produced could be considered lacking in diligence and competence.

Nevertheless, the intersection in skillset between data science and legal analysis is still nascent and it is unrealistic to expect all legally trained personnel to be trained in data science to the extent required to interpret the predictions of machine learning models without aid. Explainable AI (XAI) techniques and research have been rising in popularity since 2020 (Linardatos, Papastefanopoulos, and Kotsiantis, 2020) but have not been adopted in legal tech. XAI refers to methods that attempt to help the end user of a model understand how it arrives at its result (Danilevsky et al., 2020). XAI for NLP has been designed to combat issues of opacity by reducing the abstraction of NLP techniques so that end users can understand how the model arrived at a decision. XAI can therefore bridge the gap between the lawyer and the data scientist by using XAI techniques to explain the predictions of legal tech AI models.

However, there is not much consensus about what "explainable" means. While the general agreed upon goal of XAI is to "completely, accurately and clearly quantify the [model's] logic", there is no consistent use of the terms "explainability", "interpretability" and "transparency" (Danilevsky et al., 2020). Interpretability is sometimes used to describe the model's internal logic and explainability as the ability of the user to understand that logic. Explainability is also sometimes described as techniques to explain the model's logic post-hoc without necessarily being representative of the model's true decision (Rosenfeld, 2021). Since a debate on the "correct" definition of explainability is out of the scope of this capstone, I use explainability, interpretability and transparency interchangeably.

Another area of XAI is differentiating what explainability means to different users of the model who have different purposes for the explanations. For example, what could be explainable to data scientists may not be explainable to laypersons. While data scientists may find that more technical details about the model would make the model more explainable, laypersons might be more confused if too many details are provided (Rosenfeld, 2021). Therefore, assessing the effectiveness of XAI is highly dependent on the specific context and needs of the users.

In this capstone, XAI is assessed within the context of data privacy. The widespread collection and use of data by organisations in recent years has led to an increase of regulations governing data privacy, with 145 countries having enacted data protection legislation in 2021 (Gstrein and Beaulieu, 2022). With more sophisticated regulation comes increased difficulties for organisations to ensure that they are complying with these regulations, and for governments to enforce them. Given that explainability is context and user dependent, the below analysis explains why XAI is important to the three different stakeholders in data privacy: the consumer, organisation and state.

For the consumer, it is uncontroversial that data privacy policies on websites and software are rarely read, and even if they are read, consumers are unlikely to fully understand them because of the use of extensive legalese. The ease of reading data privacy policies was measured to be roughly the same as compared to academic articles such as the Harvard Law Review. The average policy takes 17 minutes to read, and the annual reading time per consumer is more than 400 hours which is more

than an hour per day[3] (Wagner, 2022). This increasing unreadability of data privacy policies poses a significant challenge to the effectiveness of data privacy regulation given that the current model of regulation depends on the consumer giving consent that is informed and freely given (Mantelero, 2014). Combined with the increasing collection and use of data by organisations, consumers have diminishing control over how their data is collected and used. Therefore, XAI could be helpful to consumers since XAI can be used to automate legal analysis of data privacy policies and explain the implications of the analysis in simple terms.

Organisations in the EU are subjected to a data subject's "right to explanation" under the General Data Protection Regulation (GDPR). Data subjects have the right to obtain an explanation of a decision reached solely through automated processing[4]. For example, a bank could use AI to predict the probability of a customer defaulting on a loan. This prediction could be used to justify a decision to deny a loan to the customer. Under the GDPR, the customer has the right to not be subjected to such automated processing and obtain an explanation as to why the model made such a prediction. Therefore, the use of opaque AI could potentially affect the organisation meeting their legal obligations.

In Singapore, as part of Personal Data Protection Commission's (PDPC) guidance on the operations management of AI models, organisations are advised to provide explanations on how AI models are incorporated into

[3] Assuming that a consumer visits 1462 unique websites each year.

[4] This is a plausible interpretation when reading Recital 71 in conjunction with Art 22 of the GDPR.

the decision making process of the organisations so as to build understanding and trust with those stakeholders that use their products (Personal Data Protection Commission, 2020). In terms of communications with their stakeholders, organisations are advised to develop policies on the type of explanations and when to provide them. Such communication could include explanations as to how the AI model was used in a specific decision.

Despite debate about the scope of the right to explanation and whether it is binding, (Chesterman, 2021a), such legislation in addition to the PDPC's guidance for explainability supports the need for further development of XAI specifically in the context of data privacy. The models in this capstone can support the adoption of AI in organisations by providing explanations so as to aid the organisations in checking whether their privacy policies are in compliance with data privacy regulations.

One of the state's concerns when AI is used in legal decision making is the integrity of the judicial system and regulatory activities (Chesterman, 2021b) since the legal system depends as much on the justification of its decisions as it does on the decisions themselves. In the context of Singapore's Personal Data Protection Act (PDPA), there is an appeal process for cases appearing before the PDPC and the higher courts can overturn the PDPC's decisions[5]. However, when overturning the PDPC's decisions, the higher courts do not disagree with the outcome but they disagree with the reasoning of the PDPC's decision. Assuming that AI models are sophisticated enough to write (or at least assist the writing) court judgements, if there was no explanation of the model's writing, there would be

[5]Per s48Q, s48R and s54 of the PDPA.

little basis for the higher courts to overturn judgements made by lower courts.

Hence, the importance of XAI applies across all three stakeholders in data privacy. Consumers can benefit from XAI by making privacy policies more readable. Organisations can use XAI to be more accountable to consumers and regulators by providing explanations for automated decisions. Transparency brought by XAI also helps the state to regulate organisations' use of AI.

1.2 Research questions and roadmap

Therefore, I focus on NLP and XAI in the specific context of data privacy, as this context provides a realistic and practical scope for evaluating the explainability of AI models[6]. To this end, I have four research questions:

1. Which AI classifiers perform the best on a data privacy dataset in terms of traditional performance metrics?

2. Which classifiers are the most explainable to users that include laypersons and users with domain knowledge in data science and the law?

3. How would users rate the explainability of a selected XAI technique?

4. What are the differences in explainability if users were asked to consider the predictions of these classifiers from the perspective of a

[6]All code and analysis used in this capstone can be found at https://github.com/TristanKoh/capstone-repo/.

consumer, an organisation or the PDPC (as the regulatory authority for data privacy)?

To investigate these questions, I train and evaluate the performance of AI models that are able to identify data privacy practices in apps' data privacy policies. A privacy practice describes a certain behaviour of an app that can have privacy implications[7]. XAI methods are then used to visualise why the models made such predictions, and are assessed for effectiveness by conducting a survey asking respondents to rate these visualisations according to certain metrics related to explainability.

This capstone's objectives come broadly within the evaluation of XAI techniques. While there are studies that evaluate XAI and non-empirical research that investigates the relevant standard for explainability across different areas of law, I adopt an uniquely empirical methodology to assess XAI within a data privacy context across the different use cases of stakeholders. The closest study using similar methodology is one that asked lawyers to rate visualisations generated by different XAI methods (Górski and Ramakrishna, 2021). However, this study was conducted within the context of the US Board of Veterans' Appeals, and did not consider evaluating explainability from different use cases.

Chapter 2 introduces the dataset and provides a literature review and methodology of the selected NLP models, XAI techniques, and methods used to assess XAI. Chapter 3 presents an exploratory data analysis of the dataset. Chapter 4 reports the performance of the classifiers that were trained on the dataset. Chapter 5 reports and discusses the results of the

[7]This will be further elaborated on when the dataset is introduced.

survey and Chapter 6 closes the capstone with limitations and a general discussion of the findings in relation to the research questions.

2 Literature review and methodology

There are four major components to this capstone: The dataset, model training, application of XAI techniques to the trained models and evaluating the effectiveness of these XAI techniques. I provide a literature review and describe the chosen methodology of these components below.

2.1 Dataset: The APP-350 Corpus

The APP-350 Corpus consists of 350 annotated Android app privacy policies. Each data practice ("practice") consists of a data type and a modality. A data type describes a certain behaviour of an app that can have privacy implications (e.g. collection of a phone's device identifier or sharing of its location with ad networks). There are two modalities: `PERFORMED` (i.e. a practice is explicitly described as being performed) and `NOT_PERFORMED` (i.e. a practice is explicitly described as not being performed). Altogether, 57 different practices were annotated. As not all practices had sufficient numbers to train models of sufficient baseline performance, I focused on

Data Type	Description
Contact_E_Mail_Address	The policy describes collection of the user's e-mail.
Identifier_Cookie_or_similar_Tech	The policy describes collection of the user's HTTP cookies, flash cookies, pixel tags, or similar identifiers.
Identifier_IP_Address	The policy describes collection of the user's IP address.
Location	The policy describes collection of the user's unspecified location data.

TABLE 2.1: List of top 5 data practices and their descriptions.

Party	Description
1stParty	The policy describes collection of data from the user by the app publisher.
3rdParty	The policy describes collection of data from the user by ad networks, analytics services, or other third parties.

TABLE 2.2: List of modalities.

training models on the top 5 practices by frequency[1]. The data types and modalities that are used for the rest of the capstone are provided in Table 2.1 and 2.2.

The APP-350 Corpus was annotated and used in a broader project to train machine learning models to conduct a privacy census of 1,035,853 Android apps (Zimmeck et al., 2019). The authors downloaded the data privacy practices of all apps from the Play Store with more than 350 million installs (which totalled 247 apps) and 103 randomly selected apps with 5 million installs. In total, the researchers collected the data privacy policies of 350 apps. All 350 policies were annotated by one of the authors, a lawyer with experience in data privacy law. To ensure reliability of annotations, 2 other law students were hired to double annotate 10% of the corpus. With a mean of Krippendorff's $\alpha = 0.78$[2], the agreement

[1]The performance of the models of the top n frequently occurring practices is later elaborated on in Chapter 4.

[2]Krippendorff's α is a measure of agreement, with $\alpha > 0.8$ indicating good agreement, $0.67 <= \alpha <= 0.8$ indicating fair agreement, and $\alpha < 0.67$ indicating doubtful agreement.

between the annotations exceeded previous similar research.

Since the focus of this capstone is to assess the explainability of XAI methods specifically within a data privacy context, this dataset was chosen as it contains real-world data privacy practices as these policies were scraped from Google PlayStore apps. Thus training XAI models on such a dataset would provide a realistic insight as to the performance of these models. APP-350 is also a labelled dataset, allowing easy validation of results. If an unlabelled dataset was used, unsupervised training would have to be conducted. The performance of such models would likely be much lower because NLP models for specific vocabulary like law are still not as sophisticated as models trained on general vocabulary. Further, there are few data privacy specific labelled datasets to begin with.

2.1.1 Data pre-processing

The annotated privacy policies were originally in `.yml` format, with one `.yml` file containing one app data privacy policy. As explained above, each data privacy policy is labelled at both the sentence and segment level. The data was restructured from `.yml` to `.csv`, with one `.csv` file containing annotated sentences and the other containing annotated segments. By having two levels of text data for model training, this would provide another dimension to compare model performance on.

2.2 NLP models and text representations

2.2.1 Literature review

In NLP, text representations are methods that translate plain text to mathematical representations that can be understood by a computer. The easiest and most intuitive text representation is a purely count-based approach, where each word is represented by their frequency in the sentence. However, this does not capture the semantic and syntactic differences of words (Liu et al., 2020). More advanced text representations such as GloVe have been created that are better able to capture these differences, as seen in how GPT-4 has passed the bar exam. However, to capture such semantic meaning requires more abstraction and further increases the opacity and decreases the interpretability of NLP models. Hence, there is an inverse relationship between performance / opacity and interpretability. This is typically described as the "black-box" problem of AI: only the inputs and outputs to the system can be observed, but how the model derived the outputs from the inputs is not known (or at least not easily understood) because it is difficult to know exactly how the model is programmed (Zednik, 2021).

Engineering tasks for legal NLP can be categorised according to tasks such as classification (i.e. such as classifying the data practice of a sentence, which is what this capstone is about) and text generation (i.e. generating legal documents). In terms of models and word representations used for these tasks, there seems to have been two waves of models that are based off neural networks: the first includes word2vec and doc2vec which were introduced around 2014, while the second includes ELMo

(Embeddings from Language Models), BERT (Bidirectional Encoder Representations from Transformers) and GPT that were introduced around and after 2017 (Katz et al., 2023b). The latter more advanced models are first trained on the text corpus to generate text representations and then fine tuned for a specific task, such as classification. As alluded to earlier, the main challenge facing NLP is capturing the contextual and semantic meaning of language in a machine-understandable way. These latter models use slightly different architectures, training data and fine tuning to capture more contextual information from the text corpus.

However, while these neural networks are undoubtedly more sophisticated, if simpler models are able to produce reasonable performance, these models should be used instead. Much depends on the complexity of the text corpus, and the engineering task that the models are made for. In the case of the APP-350 corpus, Zimmeck et al. were able to use classical machine learning models instead of neural networks to achieve reasonable performance to classify sentences according to data practices. Neural networks are contrasted with classical machine learning models that include logistic regression, decision trees, and support vector machines. Classical models are much less complex and therefore more explainable than neural networks, and require much less data to train. Given that the focus is on explainability, and in any case the APP-350 corpus is probably too limited to train neural networks[3], I focus only on training classical models for prediction in this capstone.

Aside from classical models for prediction, there is also another category of "classical" text representations. These are those that fall under

[3]For example, BERT was trained on the whole Wikipedia (and more), while GPT-2, the second iteration of GPT, was trained on 45 terabytes of data.

sparse vector representation (e.g. Bag-of-Words, Tf-IDF), as compared to the more contemporary and dense vector representation mentioned above that are produced using neural networks (e.g. word2vec, BERT or GPT). These sparse vector representations are purely count based and use one-hot encoding to convert a sentence into a vector, similar to the example mentioned in Section **??**. For dense vector representations, unlike models used for engineering tasks which have to be trained specifically for the task, it is possible to use pre-trained vectors, and then feed these as features into a classical model. Being pre-trained, this overcomes the limitations of the dataset and computing resources. However, explainability could still be affected compared to sparse vector representation.

For the models trained on the APP-350 corpus, Zimmeck et al. used a union of Tf-IDF vectors and manually crafted features which were Boolean values indicating the presence or absence of frequently occurring keywords that occurred in each data practice. They also conducted an optional pre-processing sentence filtering step which removes a segment's sentence from further processing if the sentence does not contain keywords associated with the data practice. This improved the performance of about half their classifiers. In terms of tokenisation, they lowercased all characters, removed non-ASCII characters, did not conduct stemming, normalised whitespace and punctuation, and created unigrams and bigrams. For prediction, individual classifiers were then trained for every data practice. For all the classifications (except for four categories), they trained SVC (support vector classification) models using the scikit-learn's implementation with a linear kernel, with five-fold cross validation. For

the four policy classifications, word-based rule classifiers were used instead because of the limited number of training data.

For the top 5 data practices that this capstone is concerned about, Zimmeck et al. were able to achieve F1 scores ranging from 77% (Identifier Cookie 1st Party), 91% (Contact Email Address 1st Party) to around 90% for location related data practices[4].

2.2.2 Methodology

Since classical models have performed well for the APP-350 corpus as seen above, I use Zimmeck et al.'s model choice and text representation of Tf-IDF as a guide for this capstone. However, for the sake of evaluating different types of word representations, I also use a pre-trained dense vector representation, GloVe (Global Vectors).

Text representations

I use Tf-IDF (term frequency - inverse document frequency) and GloVe as the two text representations. The idea behind Tf-IDF is that words that are frequent in a document but rare across the rest of the corpus are more important in distinguishing the document from others. The Tf-IDF metric for a word in a document is calculated by multiplying two different metrics:

1. Term frequency (TF) of a word in a document. This is the number of times the word appears in a document.

[4]Not all the data practices were used in training ($n = 188$) and testing ($n = 100$) these classifiers, so there is no reference performance for some of the data practices that are used in this capstone.

2. Inverse document frequency (IDF) of the word across a set of documents. This is calculated by taking the total number of documents and dividing it by the number of documents that contain the specific word. This calculates the rarity of the term across all the documents. The closer the IDF of a word is to 0, the more common the word is.

Mathematically, the Tf-IDF score for the word t in the document d from the document set D can be stated as such:

$$tf - idf(t, d, D) = tf(t, d) \cdot idf(t, D)$$

where

$$tf(t, d) = log(1 + freq(t, d))$$
$$idf(t, D) = log\left(\frac{N}{count(d \in D : t \in d)}\right)$$

While Tf-IDF is easy to calculate, one of its limitations is that it is a purely count-based metric. Tf-IDF does not take into account the context of the word. For example, Tf-IDF does not capture the semantic relationship between words such as the difference between King and Queen, vs King and Pauper. In this regard, dense vector representations are able to better model such semantic differences mathematically in the following way: King $-$ Man $+$ Woman $=$ Queen (Vylomova et al., 2015). Hence, to compare whether a more sophisticated text representation is more explainable, I chose to use pre-trained GloVe word embeddings as well. For convenience, GloVe was chosen since the pre-trained vectors are freely available for use.

Strictly speaking, GloVe is an unsupervised learning algorithm for obtaining vector representations for words. The length-50 vectors used in this capstone are pre-trained by Stanford researchers using this algorithm on Wikipedia and news articles (Pennington, Socher, and Manning, 2014). While the specifics of GloVe are out of scope for this capstone, the main idea behind GloVe is to learn text representations by considering the statistics of words that occur together in a corpus. These statistics captures how frequently different pairs of words occur together in the text. GloVe builds a matrix from these statistics, where each element represents the number of times two words appear together in the same context. The matrix is then factorized into a product of two low-rank matrices, where each row in one of the matrices represents a vector representation of a word. This results in dense, fixed-size vector representations for each word in the corpus that encode their meaning and relationships with other words in the corpus.

Model choice

As Zimmeck et al. found that LinearSVC produces the best performance, I use SVC with a linear kernel as well. I also use scikitlearn's logistic regression as a baseline classifier for its simplicity. While I also trained ensemble classifiers (AdaBoost, GradientBoost, Random Forest), the performance of these models were roughly the same or less than logistic regression and SVC. Hence, I excluded them from the rest of the capstone.

2.3 XAI methods for NLP

2.3.1 Literature review

XAI for NLP have two broad characteristics: Local vs global, and self-explaining vs post-hoc. Thus, there are four possible categories of XAI as seen in Table 2.3.

Category	**Description**
Local Post-Hoc	Explain a single prediction by performing additional operations (after the model has made a prediction)
Local Self-Explaining	Explain a single prediction using the model itself (calculated from information made available from the model as part of making the prediction)
Global Post-Hoc	Perform additional operations to explain the entire model's predictive reasoning
Global Self-Explaining	Use the predictive model itself to explain the entire model's predictive reasoning

TABLE 2.3: The four categories of XAI for NLP, adapted from Danilevsky et al., 2020.

Some classical models such as logistic regression can be considered local and global self-explaining models. The coefficients of the features show the relative importance of each feature towards making the prediction. These coefficients are generated as part of the model training process and make it self-explaining. Logistic regression is also globally explainable since it models a deterministic function (the logit), and this function serves as the "reason" for any prediction by the model. However, neural networks are not self-explaining by nature, in part as they are non-linear and are fitted onto high-dimensional data, in additional to the large numbers of parameters that make it difficult to ascertain the interaction between different features when a prediction is made. Therefore, post-hoc methods of explanability have to be used instead.

One such example is LIME (Local Interpretable Model-agonistic Explanations) (Ribeiro, Singh, and Guestrin, 2016b). LIME is a local post-hoc XAI method that is also model agnostic. Given a trained NLP model, LIME generates a set of perturbations on the input sentence by randomly replacing some of the words with similar words. These perturbed samples are passed to the NLP model for prediction. Using these predictions, LIME fits a simple self-explaining model (such as a linear regression or a decision tree) to explain these predictions. Through this local self-explaining model, LIME is able to generate the contribution of each feature to a particular prediction for a single sample. Figure 2.1 is an example of how an explanation of LIME can be visualised. Since LIME is a local, post-hoc method that uses a surrogate model to generate explanations, one disadvantage is that it may not fully represent how the underlying model actually made its prediction. Nevertheless, LIME has been found to be able to generate explanations that are faithful to the underlying model even just by using linear models for the surrogate model (Ribeiro, Singh, and Guestrin, 2016b).

FIGURE 2.1: Sample visualisation of LIME

Explainability techniques should also be distinguished from visualisation techniques (Danilevsky et al., 2020). Explainability techniques are ways to generate the raw mathematical justifications that led to the final explanation presented to the end users. For example, feature importance

is one technique that identifies the most important words / phrases in the sentence that led to the prediction, such as LIME. In contrast, visualisation techniques are different ways to present these mathematical justifications to the end user that can deal with how the explanations can be aesthetically presented. One example is a saliency heatmap (such as Figure 2.1), which highlights the combination of words / phrases that gave rise to the prediction at differing intensities. In this regard, there are also different Python visualisations of LIME. This difference between explainability and visualisation is akin to the difference between proportions and pie charts. While proportions can be presented in pie charts, there is no necessity for this to be so. Proportions can also be presented in a bar chart. Hence, these are two related, but also separate areas of research that go towards the overall objective of explainability.

2.3.2 Methodology

I chose to use LIME because it is model agnostic which allows easy experimentation with different combinations of word representations and models without needing much tweaking. It is also one of the few XAI packages that are easily implementable in Python. In particular, I use the implementation by the LIME creators (Ribeiro, Singh, and Guestrin, 2016a). I apply LIME on the different combinations of word representations and models and compare the explainability of LIME's explanations. I presented these explanations to the survey respondents without any amendments to the code output, except for minor editing to the labels in the figure because some labels were cut-off.

2.4 Evaluating the effectiveness of XAI methods

2.4.1 Literature review

There is not much general consensus as to the methods and standards of what makes a particular XAI technique effective. This is because explainability is highly context dependent, as alluded to in Section **??**. In terms XAI and the law, the main challenge towards greater adoption of XAI is determining what exactly would be considered sufficiently explainable for legal purposes. While post-hoc methods have been argued to be insufficient to pass muster for non-discrimination law (Vale, El-Sharif, and Ali, 2022), it should be noted that legal standards for explainability differ too. What could be insufficiently explainable for non-discrimination law might not apply *mutatis mutandis* to data privacy law under the GDPR (for instance). Indeed, other authors have argued that tort and contract law in medical diagnostics may sometimes mandate using a higher performing model over a more explainable model depending on the specific use case (Hacker et al., 2020). In any case, this capstone makes no claim as to what the applicable standard for explainability is for data privacy, and merely serves as empirical insight as to the perceptions of explainability by potential data subjects.

Apart from the specific requirements of the law, the general field of evaluating XAI is also nascent. Three categories of evaluating XAI have been proposed (Danilevsky et al., 2020): First is informal examination of explanation, which are high level discussions of how generated explanations cohere with human intuition. Second is comparison to ground truth, such as quantitative metrics like precision / recall / F1. Third is

human evaluation, which is simply to ask humans to evaluate the effectiveness of the explanations. There are a few methods to conduct human evaluation of XAI as covered in Table 2.4. Apart from how XAI is evaluated, what is evaluated includes fidelity (how much the explanation reflects the actual workings of the model), comprehensibility (how easily understood are the explanations), and others. Many studies also do not state what is being evaluated in XAI assessments.

Method	**Description**
Binary forced choice	Humans are presented with pairs of explanation, and must choose the one that they find of higher quality.
Forward simulation / prediction	Humans are presented with an explanation and an input, and must correctly simulate the model's output.
Counterfactual simulation	Humans are presented with an explanation, an input, and an output, and are asked what must be changed to change the model's prediction to a desired output.

TABLE 2.4: Methods of measuring human evaluation of XAI. Adapted from (Doshi-Velez and Kim, 2017)

Human evaluation has been conducted within a legal context (Górski and Ramakrishna, 2021). The authors trained a CNN on a dataset that contained 50 decisions issued by the Board of Veterans' Appeals of the US Department of Veterans Affairs. Each sentence was annotated according to a type of legal reasoning (e.g. fact finding, legal reasoning, application of legal rule etc.). The focus of the research was to evaluate different XAI methods and lawyers' perceptions and expectations of XAI. Lawyers were asked to rate different XAI visualisations of the same 6 correctly classified sentences on a scale from 0 (worst) to 10 (best). Further, the lawyers that were surveyed were generally optimistic about the use of AI in law, with most agreeing that the explanations generated should be supplemented by further background knowledge, and that the model

should be fair, lack bias, transparent and have awareness of the consequences of the AI-assisted / AI-automated decision. Lawyers also agreed that AI should increase automation and reduce their manual effort.

Further, the purpose of generating the explanation is important towards assessing whether the explanation is effective. Here are some factors which purposes can be categorised into:

1. *Global vs local:* A scientific understanding of a model predicting rainfall would require global explanation since a scientist would be concerned with finding long term correlations, while a consumer who failed to obtain a loan would only require a local explanation to justify why the model made such a prediction.

2. *Area and severity of incompleteness of the problem:* For example, a potential buyer of an autonomous vehicle (AV) would have a general curiosity about how they work which requires more inquiry to figure out what aspect of the model to focus on explaining, while a regulator may have specific explanation requirements such as whether the model detects a pedestrian 10 metres before collision.

3. *Time constraints:* Consumers that get a bank loan rejected would likely have more time to scrutinise the explanation compared to a compliance manager in the bank who has to process hundreds of claims a day.

4. *Nature of user expertise:* A layperson with no training in credit worthiness would not expect details relating to large, complicated models, compared to an auditor that would expect such models to accurately model how human evaluations are made.

In addition, as mentioned earlier, there is no agreed upon definition of "explainability". Explainability could be influenced by other value-laden terms, such as, *inter alia*, trust, effectiveness, fairness, bias, and the consequence of making a wrong prediction (Rosenfeld, 2021). The relative importance of these values and overall requirement of explainability is necessarily dependent on the user of the explanation. All these values are particular importance to a data privacy context, since data privacy operates within the broader legal system. The notion of "rule of law" encompasses these values and legal systems are judged according to them as well (Greenstein, 2022). However, currently, there is no definitive position as to how these values which are interlinked with explainability can be translated from legal norms to XAI techniques.

2.4.2 Methodology

With the foregoing discussion in mind, I expand upon and tweak the methodology of Gorski and Ramakrishna. The primary method of evaluating LIME in this capstone is human evaluation through a survey. The survey was designed containing the following parts[5]:

1. **Demographic data and beliefs relating to data privacy and AI.** Respondents were asked to provide their major or subject area expertise if they had graduated. This is used as a proxy for their level of expertise in AI or data science. Respondents are also asked to rate how much they think AI is useful, is a risk as well as how far they are concerned about their data privacy. This is used to understand

[5]A copy of the full survey is included in Appendix A.

the underlying values the respondents' have such that their evaluation of LIME and the models can be evaluating in the context of these values. Further cross-sectional analysis of the survey data can also be conducted.

2. **Perceptions of explainability from the perspectives of three stakeholders in data privacy and introduction to the dataset.** After a brief explanation of the data practices and how NLP works, I provide three contexts relating to data privacy to measure the variability of explainability according to different use cases and levels of data science & law expertise (app developer, PDPC & user). After displaying each context, the respondents are required to rate their perception of explainability in that particular context across four metrics: Effectiveness of model, Fairness, Risk to society, and trust in the model. These metrics are to measure how respondents' values that relate to explainability can vary according to the context and the purpose of generating the explanation.

3. **Rating of LIME visualisations of four classifications by the same model (Logistic regression + Tf-IDF).** Respondents were asked to rate how far they understood why the model made the prediction, and how far they found the visualisation easy to interpret. The intent is to evaluate both the underlying XAI technique (LIME) and the visualisation of the technique. For three classifications, respondents were given a counterfactual sentence that replaced words that were considered as important by the model, and asked to predict whether the sentence would be classified under `Identifier Cookie`

`1st Party`. Predicting counterfactuals is meant to provide another metric to measure explainability, aside from self-reported scores.

4. **Choosing the more interpretable visualisation from a pair of identically classified sentences.** This is to evaluate which text representation and model are more explainable. Part 4 presents 3 pairs of visualisations from SVC + GloVe and Logistic regression + GloVe (controlling for text representation) and Part 5 presents another 3 pairs of visualisations from SVC + Tf-IDF and SVC + GloVe (controlling for model). The respondents' choices are then totalled up and compared against traditional performance metrics (P/R/F1) of the classifiers, to compare whether there is a relation between explainability and model performance.

5. **Repeating the questions for the same three contexts.** Respondents are asked to view the same three contexts and questions. This is to evaluate whether viewing the explanations significantly changed how the respondents rated the explainability of the visualisations, and whether there are relations and trends of the individual metrics.

The survey was hosted on Google Forms, and respondents were recruited through friends and acquaintances of this author. There was no restriction on who could participate, other than the age requirement of 18 and over for NUS / Yale-NUS students, or 21 and over for the general public due to Yale-NUS Ethics requirements.

3 Exploratory Data Analysis (EDA)

As mentioned above, there are two levels of text data: Sentence level and segment level. As I train and compare the performance of models on both levels of data, I also conducted the EDA on both levels.

3.1 Sentence level

There are a total of 18829 annotated sentences. Table 3.1 and 3.2 show some summary statistics of the top 10 and bottom 10 frequently occurring practices at the sentence level.

practice	counts	sentence_length_mean	sentence_length_median	counts_percentage
Identifier_Cookie_or_similar_Tech_1stParty	2107	25.389654	22.0	11.2%
Contact_E_Mail_Address_1stParty	2106	28.651472	25.0	11.2%
Location_1stParty	1514	29.159181	24.0	8.1%
Identifier_Cookie_or_similar_Tech_3rdParty	1250	27.318400	24.0	6.6%
Identifier_IP_Address_1stParty	1005	30.913433	27.0	5.3%
Contact_Phone_Number_1stParty	970	29.117526	25.0	5.2%
Identifier_Device_ID_1stParty	697	32.377331	28.0	3.7%
Contact_Postal_Address_1stParty	597	28.907873	26.0	3.2%
SSO	504	32.565476	28.0	2.7%
Demographic_Age_1stParty	428	33.074766	26.0	2.3%

TABLE 3.1: Summary statistics for top 10 occurring practices at sentence level.

In total, the top 10 frequently occurring practices make up approximately 60% of the dataset. The bottom 10 frequently occurring practices make up approximately 1% of the dataset.

practice	counts	sentence_length_mean	sentence_length_median	counts_percentage
Identifier_Mobile_Carrier_3rdParty	35	47.057143	30.0	0.19%
Contact_ZIP_3rdParty	34	40.176471	41.0	0.18%
Identifier_SSID_BSSID_1stParty	33	28.060606	24.0	0.18%
Contact_Password_3rdParty	33	24.181818	20.0	0.18%
Contact_City_3rdParty	24	18.000000	14.0	0.13%
Contact_Address_Book_3rdParty	17	39.647059	34.0	0.1%
Identifier_IMSI_1stParty	13	48.153846	44.0	0.07%
Identifier_SIM_Serial_3rdParty	5	41.200000	54.0	0.03%
Identifier_IMSI_3rdParty	4	54.250000	47.5	0.02%
Identifier_SSID_BSSID_3rdParty	2	65.500000	65.5	0.01%

TABLE 3.2: Summary statistics for bottom 10 frequently occurring practices.

According to Table 3.3, there does not seem to be much variation in sentence length for the top 10 frequently occurring practices, since the standard deviation for the mean is approximately 2.5 words and the median 1.9 words. This could indicate similar sentence complexity across the practices.

	sentence_length_mean	sentence_length_median
count	10.000000	10.000000
mean	29.747511	25.500000
std	2.468191	1.900292
min	25.389654	22.000000
25%	28.715572	24.250000
50%	29.138353	25.500000
75%	32.011357	26.750000
max	33.074766	28.000000

TABLE 3.3: Summary sentence statistics for top 10 frequently occurring practices.

3.2 Segment level

There are in total 21623 segments. However, there are 11422 segments without any annotated practice. Hence there are 10201 annotated segments. Table 3.4 and 3.5 shows some summary statistics of the top 10

and bottom 10 frequently occurring practices at the segment level.

practice	counts	segment_length_mean	segment_length_median	counts_percentage
Contact_E_Mail_Address_1stParty	1105	84.118552	72.0	10.83
Identifier_Cookie_or_similar_Tech_1stParty	858	81.913753	68.5	8.41
Location_1stParty	821	89.794153	70.0	8.05
Identifier_IP_Address_1stParty	590	96.984746	73.0	5.78
Contact_Phone_Number_1stParty	565	90.371681	67.0	5.54
Identifier_Cookie_or_similar_Tech_3rdParty	524	100.339695	86.5	5.14
Identifier_Device_ID_1stParty	446	96.704036	72.0	4.37
Contact_Postal_Address_1stParty	364	85.598901	69.5	3.57
SSO	274	99.463504	86.5	2.69
Demographic_Age_1stParty	259	96.200772	80.0	2.54

TABLE 3.4: Summary statistics of top 10 frequently occurring practices by segment.

practice	counts	segment_length_mean	segment_length_median	counts_percentage
Identifier_IMEI_3rdParty	23	115.347826	93.0	0.23
Identifier_Mobile_Carrier_3rdParty	21	142.000000	130.0	0.21
Contact_Password_3rdParty	18	70.555556	65.0	0.18
Identifier_SSID_BSSID_1stParty	16	86.062500	71.0	0.16
Contact_Address_Book_3rdParty	14	296.928571	78.5	0.14
Identifier_IMSI_1stParty	11	92.363636	78.0	0.11
Contact_City_3rdParty	8	100.500000	104.0	0.08
Identifier_SIM_Serial_3rdParty	3	85.333333	65.0	0.03
Identifier_IMSI_3rdParty	3	62.333333	65.0	0.03
Identifier_SSID_BSSID_3rdParty	2	105.500000	105.5	0.02

TABLE 3.5: Summary statistics for bottom 10 frequently occurring practices.

The top 10 practices make up approximately 57% of the dataset. The bottom 10 practices make up approximately 1.2% of the dataset.

According to Table 3.4 , there does not seem to be much variation in sentence length for the top 10 frequently occurring practices, since the standard deviation for the mean is approximately 6.7 words and the median 7.2 words. This could indicate similar segment complexity across the practices.

	segment_length_mean	segment_length_median
count	10.000000	10.000000
mean	92.148979	74.500000
std	6.683275	7.230337
min	81.913753	67.000000
25%	86.647714	69.625000
50%	93.286227	72.000000
75%	96.914568	78.250000
max	100.339695	86.500000

TABLE 3.6: Summary segment statistics for top 10 frequently occurring practices.

4 Performance of models

4.1 Models used and classification metrics

I use Logistic Regression, SGDClassifier and SVC classifiers and compare the weighted precision, recall and F1 scores. Precision, Recall and F1 scores different metrics are used to assess the performance of classifiers. They are stated mathematically below.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$
$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$
$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision + Recall}}$$

Since there is neither a high risk of identifying false positives or false negatives, F1 score is primarily used to assess the performance for the models. As the distributions of records across the practices are uneven, I focus on the weighted average of F1 scores as the most important indicator of model performance.

4.2 Performance of classifiers for top N practices

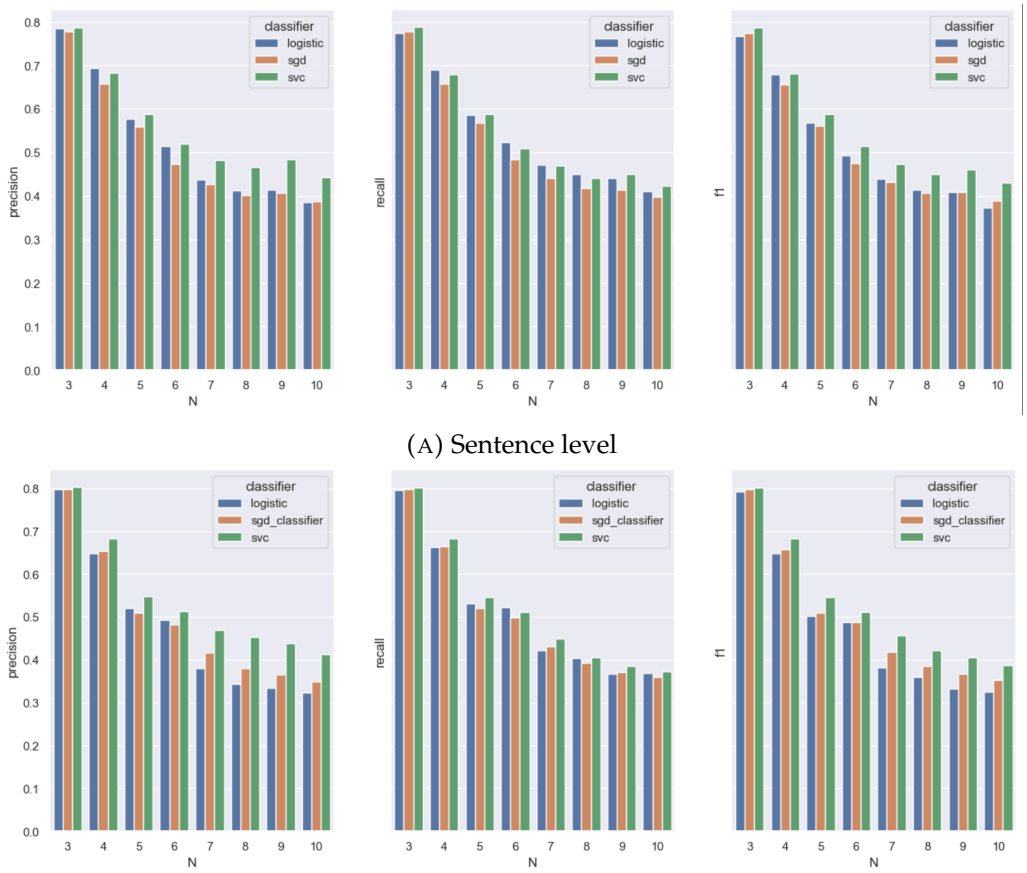
As seen in Table 3.1 and Table 3.4 above, there is an uneven distribution of records across the practices. Further, the bottom 10 frequently occurring

practices for both sentence and segment level only contains about 2 to 35 records for each practice. Given that there are in total 57 practices for the entire dataset, and there is not a uniform distribution of occurrences. Training models on all 57 practices would likely lead to low performance since there are not enough records for all 57 practices. Thus, to find an optimal balance between model performance and still maintain a realistic sample of practices that could appear in a real world dataset, I chose to assess model performance by first assessing the performance of the three classifiers paired with Tf-IDF for the top N (where $3 \leq N \leq 10$) frequently occurring practices at both the sentence and segment level. The metrics are in Figure 4.1.

Generally we see that the SVC classifier performs the best across the metrics and across the top N frequently occurring practices for both sentence and segment level. This corresponds with the findings by the researchers as they also found that the SVC classifier produced the best performance. Also, performance of all the models generally decreases with increasing N, which is not surprising since it gets harder for the classifier to find a decision boundary with more classes.

4.3 Performance of individual classifiers for top 5 practices

Since SGDClassifier performed similarly to the 2 other classifiers, I focus only on comparing logistic regression and SVC, with logistic regression as the baseline classifier and SVC since the APP-350 researchers used SVC. I also compare the performance of Tf-IDF and GloVe. In total, I



(A) Sentence level

(B) Segment level

FIGURE 4.1: Weighted average P/R/F1 scores for top N ($3 \leq N \leq 10$) practices

documented the performance of all 4 possible combinations of text representations and models in Tables 4.1, 4.2, 4.3 and 4.4 show the averaged metrics over a 5-fold cross validation.

Similarly to what Zimmeck et al. found, SVC + Tf-IDF provided the best overall performance with the weighted average at 66%. This is quite a significant performance difference from the lowest performing model, logistic regression + GloVe at 54%. Specifically for `Identifier Cookie 1st Party`, both logistic regression + Tf-IDF and SVC + Tf-IDF performed similarly at 68% F1, but SVC + Tf-IDF had a more balanced performance across precision and recall. Overall, these performance figures confirm

that SVC is the better performing model, though the performance difference between the benchmark logistic regression model and SVC is not that significant (from 62% to 66% weighted average F1). It is also interesting to note that GloVe embeddings performed worse compared to Tf-IDF regardless of the model used, when GloVe is the more sophisticated text representation given that only GloVe vectors capture the context of the word. Further, these results seem to correspond with a study measuring performance of supervised machine learning for classification tasks using generalised text datasets (Hsu, 2020). LinearSVC followed by logistic regression were the highest performing models, with AdaBoost and Decision Trees having comparatively lower performance.

Data Practice	**Precision**	**Recall**	**F1**	**Support**
Contact Email Address 1st Party	0.66	0.88	0.76	2106
Identifier Cookie 1st Party	0.59	0.78	0.68	2107
Identifier Cookie 3rd Party	0.64	0.35	0.45	1250
Identifier IP Address 1st Party	0.66	0.41	0.50	1005
Location 1st Party	0.69	0.49	0.57	1514
accuracy			0.64	7982
macro avg	0.65	0.58	0.59	7982
weighted avg	0.64	0.64	0.62	7982

TABLE 4.1: Logistic regression + Tf-IDF

Data Practice	**Precision**	**Recall**	**F1**	**Support**
Contact Email Address 1st Party	0.62	0.77	0.70	2106
Identifier Cookie 1st Party	0.53	0.67	0.59	2107
Identifier Cookie 3rd Party	0.50	0.34	0.41	1250
Identifier IP Address 1st Party	0.47	0.22	0.30	1005
Location 1st Party	0.52	0.48	0.50	1514
accuracy			0.55	7982
macro avg	0.53	0.50	0.50	7982
weighted avg	0.54	0.55	0.54	7982

TABLE 4.2: Logistic regression + GloVe embeddings

Data Practice	**Precision**	**Recall**	**F1**	**Support**
Contact Email Address 1st Party	0.74	0.81	0.77	2106
Identifier Cookie 1st Party	0.67	0.68	0.68	2107
Identifier Cookie 3rd Party	0.59	0.57	0.58	1250
Identifier IP Address 1st Party	0.58	0.60	0.59	1005
Location 1st Party	0.65	0.55	0.60	1514
accuracy			0.66	7982
macro avg	0.65	0.64	0.64	7982
weighted avg	0.66	0.66	0.66	7982

TABLE 4.3: SVC + Tf-IDF

Data Practice	**Precision**	**Recall**	**F1**	**Support**
Contact Email Address 1st Party	0.70	0.70	0.70	2106
Identifier Cookie 1st Party	0.61	0.51	0.55	2107
Identifier Cookie 3rd Party	0.46	0.51	0.48	1250
Identifier IP Address 1st Party	0.37	0.49	0.42	1005
Location 1st Party	0.49	0.45	0.47	1514
accuracy			0.55	7982
macro avg	0.53	0.53	0.53	7982
weighted avg	0.56	0.55	0.55	7982

TABLE 4.4: SVC + GloVe embeddings

5 Discussion of survey results

5.1 Summary of results

In total, 31 responses were collected[1]. The majority of respondents majored in law (58%) vs non-law (42%) respondents. Table 5.1 and Figure 5.1 show the demographic data and responses relating to the respondents' beliefs regarding AI and data privacy. I originally intended to conduct analysis of the results across different demographics such as those with more expertise in computer science, or those who think that decisions by AI are not as useful. However, 31 respondents are insufficient for further meaningful demographic breakdown.

Major / Expected major	**Count**	**Percentage of total (%)**
Law	18	58
MCS / Computer Science / Data Science / Statistics	3	9.7
Psychology	3	9.7
Global Affairs / Political Science	2	6.5
Environmental Studies	1	3.2
Economics	1	3.2
Life Sciences	1	3.2
Philosophy	1	3.2
Policy	1	3.2

TABLE 5.1: Demographic breakdown of respondents according to academic discipline

[1]As mentioned in Chapter 2, a copy of the survey and aggregated results can be found in Appendix A and B.

5.1.1 Part 1: Beliefs relating to AI & data privacy

Except for the questions relating to subject matter expertise (data privacy and AI), the level of agreement of law vs non-law respondents were about the same (Figure 5.1). Law respondents had less expertise in AI, while conversely, non-law respondents had less experience with data privacy. Across all respondents, while they rated that decisions by AI could be a risk to society (about 4), they also agreed that decisions by AI could be equally useful. This perhaps suggests that the respondents think the balance between "usefulness" and "risks" are not zero-sum; AI could be very helpful in solving problems, but at the same time users should be cognisant of the risks.

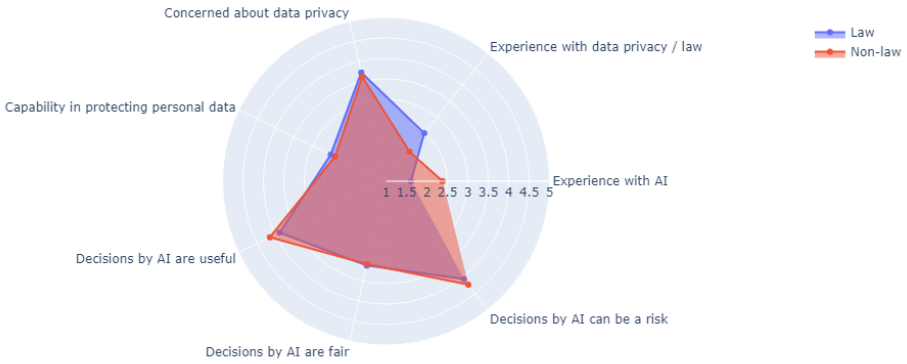
5.1.2 Part 2 & Part 6: Comparison of self-reported scores of explainability across the three contexts

Using the Wilcoxon Rank Sum Test, I tested for the following, setting $\alpha = 0.1$:

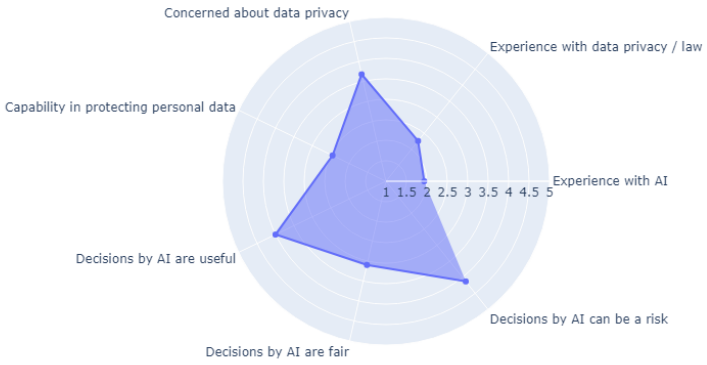
$H0$: There is no increase / decrease in scores after viewing the explainations.

$H1$: There is an increase / decrease in scores after viewing the explainations.

The 1-sided test was used to check whether the distribution underlying the difference between the initial and final paired scores was symmetric below or above 0 (SciPy, 2023). Mathematically this difference can be stated as $d = i - f$, where i and f are the scores reported before and after viewing the explanations, and d is the difference. Hence, if $d < 0$,



(A) Law vs Non-law respondents



(B) All respondents

FIGURE 5.1: Mean scores of self-reported beliefs of respondents regarding AI & data privacy. (1 = least agree, 5 = strongly agree. $n = 31$)

then $i < f$ and the scores increased after viewing. Conversely, if $d > 0$, then $i > f$ and the scores decreased after viewing.

The p-values of the tests are reported in Table 5.2 , 5.3 and 5.4. For Table 5.3a and 5.3b, I took the mean of self-reported scores across contexts, and across each metric, respectively. "Increase" and "decrease" refer

Question	Context 1: Increase	Context 1: Decrease	Context 2: Increase	Context 2: Decrease	Context 3: Increase	Context 3: Decrease
Do you think model is effective?	0.013	0.987	0.932	0.0684	0.856	0.144
Do you think model is a fair method?	0.382	0.618	0.841	0.159	0.933	0.0671
Do you think model is a risk to society?	0.756	0.244	0.428	0.572	0.825	0.175
Do you trust the prediction of the model?	0.887	0.113	0.945	0.055	0.837	0.163

TABLE 5.2: p-values comparing whether there was a statistically significant increase / decrease in the explainability scores before and after viewing explanations.

Context	p-value
1: Increase	0.369
1: Decrease	0.633
2: Increase	0.940
2: Decrease	0.060
3: Increase	0.955
3: Decrease	0.0450

(A) p-values when scores are averaged across metrics, compared by contexts

Metric	p-value
Effective: Increase	0.485
Effective: Decrease	0.515
Fair: Increase	0.826
Fair: Decrease	0.174
Risk: Increase	0.660
Risk : Decrease	0.340
Trust: Increase	0.953
Trust: Decrease	0.0468

(B) p-values when scores are averaged across contexts, compared by metrics

TABLE 5.3: p-values comparing aggregated scores by context and metric

Increase	Decrease
0.844	0.156

TABLE 5.4: p-values comparing aggregated scores across contexts and metrics

to the p-values of the 1-sided test that checks whether the scores significantly increased or decreased after viewing the explanations. For Table 5.4, the self-reported scores were averaged across both context and metric. Here are the instances when $p < 0.1$ (highlighted in red in the tables), and therefore $H0$ can be rejected in favour of $H1$ such that there was a statistically significant increase / decrease of scores after viewing the explanations:

1. Effectiveness and trust significantly decreased for context 2 (PDPC), and fairness significantly decreased for context 3 (consumer). Only effectiveness significantly increased for context 1 (app developer) (Table 5.2). While trust technically did not have a statistically significant decrease for context 1, the p-value is quite close at 0.113. Hence, there seems to be an inverse relationship between effectiveness and trust for context 1, but a positive relationship for context 2.

2. There is a statistically significant decrease in explainability metrics for context 1 and 2 when comparing the mean of the scores across the metrics (Table 5.3a) and for trust across the three contexts (Table 5.3b).

3. The overall self-reported explainability when taking the mean of the scores across both contexts and metrics did not significantly increase / decrease (Table 5.4).

Remember that the explanations given to respondents were deliberately chosen to demonstrate the limitations of the model. Hence, it can be inferred that respondents were more cognisant of such limitations when

they answered the same questions again in Part 6. Here are some inferences that can be drawn from these observations:

1. The inverse relationship between effectiveness and trust for context 1 (app developer) could be explained by the fact that respondents believed that software engineers may be more concerned with a standardised, consistent way to identify data practices, as someone who is not legally trained but is still subjected to data privacy obligations. Hence, self-reported effectiveness increased when they realised that the model would still do a better job at identifying data practices consistently compared to themselves. However, trust decreased because respondents also knew that the model could make wrong predictions. In this context, respondents were willing to bear some risk of getting predictions wrong in return for automated detection of data practices.

2. For the positive relationship between effectiveness and trust for context 2 (PDPC), respondents were more concerned about retaining discretion in their decision making process. Since they knew that the model could make wrong predictions, combined with the fact that PDPC is the regulatory body that enforces data privacy, the costs of making a wrong decision is extremely high. Hence, even though the model is more "effective" since it is automated, it is not "effective" in terms of making a legally defensible decision given that it is fallible. Trust decreased because while the respondents believed that this model could make their jobs in identifying data breaches less repetitive (in that they would not have to read every

data privacy policy manually), there is now a mismatch of expectations because respondents still have to ensure that the predictions given by the model are legally correct.

3. For the consumer, as a person who is not trained in law or data science, they are subjected to how the app developer designed the app such that the app developer failed to disclose that the app was tracking cookies. However, as consumers are not legally trained, they cannot correct this breach of their privacy rights except by relying on the fallible predictions of the model. There is a chance that the model makes the wrong predictions, and the respondent fails to make a case to the PDPC. Therefore, it is not "fair" to the consumer that though their privacy was violated, they cannot do anything about it because of their reliance on the fallible model.

4. Overall, a different weighting of these four values by the respondents can be inferred from the significance tests. The respondents were more concerned about having automated identification of data practices as a software engineer, over the risk of the model making a wrong prediction. In this case, effectiveness (in the sense of efficiency of identifying data practices) was prioritised above trust. This is contrasted with the PDPC context in which respondents were highly concerned of the risks of making a wrong decision more so than automating reading data privacy policies, and so fallible models are seen as impacting effectiveness and trust towards making sound legal decisions. The consumers are mainly concerned with their own interests and whether their rights would

be vindicated, and so fallible models would greatly hinder them from pursuing this self-interest which is unfair for them.

5. As to the comparison of aggregated scores in Table 5.3a, the fallibility of the model weighed heavily in the respondents' minds to reduce the explainability metrics in context 2 (PDPC) and context 3 (consumer). Perhaps the respondents thought that as app developers who were trained in programming, they would possess the requisite expertise to understand and work around the model's limitations compared to context 2 (PDPC) and context 3 (consumer) who may not have such technical knowledge and so are limited by the explanations that were presented to them.

6. In the case of trust being the only metric that significantly decreased across the contexts, this perhaps shows the interrelation between of model transparency and trust. Intuitively, if respondents do not understand how a process works, they are less likely to trust the results of that process. However, since the respondents were made aware of the limitations of the model, they still have reduced trust not because they did not understand the model, but that their expectations of the model in producing objective, 100% reliable predictions have been affected. Hence, there can be two plausible reasons for the loss in trust: the explanations were not effective in explaining the model, or that their expectations in the model have decreased as a result of being made aware of the limitations.

5.1.3 Part 3: Testing whether viewing more visualisations increase explainability

Analysis of the reported scores of "interpret" and "understand"

There is a decreasing trend of both understanding and interpret after viewing each explanation, though understanding has a more negative gradient than interpret (Figure 5.2). It seems that respondents found the model less generally explainable after viewing more explanations, both in terms of the method of visualising the explanation (which is the "interpret" question), as well as understanding the inner logic of the model (the "understand" question).

One interpretation of this decreasing trend suggests both the model and LIME were not very explainable. However, another interpretation could be that the respondents being more confused about how the model works globally after being exposed to specific predictions that were in themselves explainable, rather than being confused about a specific prediction of the model. As mentioned in Chapter 2, the visualisations for this part were specifically chosen to demonstrate the limits of the model by changing keywords which were strong predictors of the data practice. This means that respondents were being exposed to a more complex (and confusing) of the model globally as they found contradictions in how the model used certain keywords in some examples but not in others. Therefore, each explanation was actually effective in communicating to the respondents how that particular prediction was made, and therefore could be considered as locally explainable. However, as a whole, respondents' understandability of the model given the contradictions seen through

comparing each explanation decreased. This decrease in explainability could be likened to a variation of the Dunning-Kruger effect (Dunning, 2011), whereby the respondents who initially had no experience with the model believed that they understood the model well because they were unaware of the complexity of the model, but when they were given more information, they had more awareness of the complexity and therefore reported a drop in explainability of the model as a whole. Therefore, LIME and the model could be said to be locally explainable, but globally not explainable. If this interpretation of the results is adopted, it would not be surprising since LIME is designed as a post-hoc, local XAI technique.

Another interpretation is that the underlying explainability method also influences the respondents' perception of the explainability of the visualisation technique (or vice versa). This can be seen by how there is a positive correlation between the understand and interpret scores. This is not surprising since if the visualisation technique is unclear or poor, explainability of the model itself would also decrease since respondents' only way of viewing the results of the explainability technique is through the visualisation technique. To respondents' both the visualisation technique and the explainability technique are one and the same thing.

Analysis of predicting counterfactuals

It should be noted that the model itself gave quite ambivalent predicted probabilities for each data practice. For example, for the second counterfactual (Figure 5.4), the model predicted `Contact_Email_Address` at

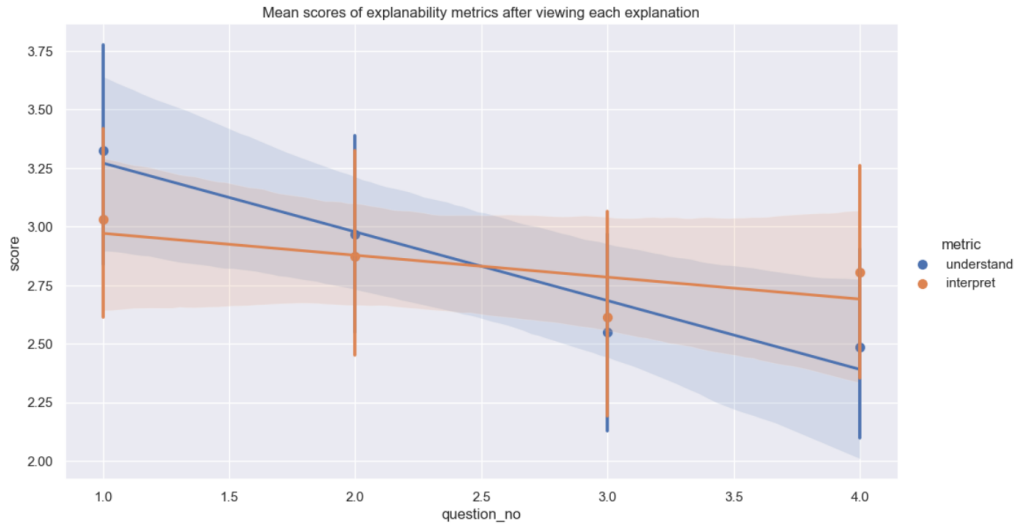
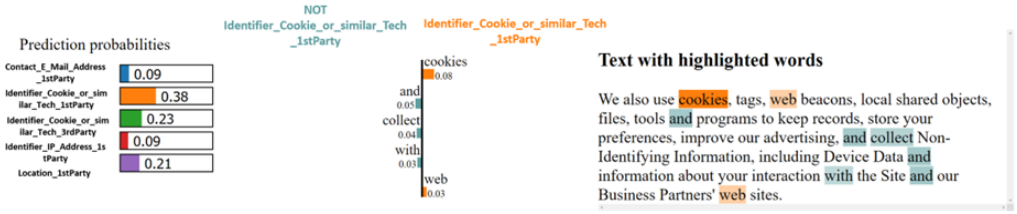


FIGURE 5.2: Trend of the mean of self-reported understanding and interpretability after viewing each explanation

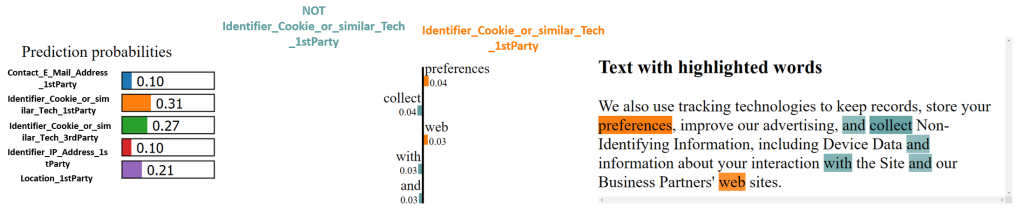
25% while `Identifier_Cookie` was predicted at 23%. Hence, the respondents' votes of predicting whether these counterfactuals would be classified as `Identifier_Cookie` should also be roughly ambivalent if the respondents actually had a good sense of how the model would predict. Only the results of the first question had some similarity to the predicted probabilities of the model, while the results of the third question differed the most from the model (Figure 5.6). Nevertheless, given the lack of any consistent trend across the questions, it is difficult to draw any definitive interpretations from the respondents' predictions of counterfactuals.

5.1.4 Part 4 & 5: Testing which model and word representation is more explainable

As there were three questions in each section to test the explainability of each pair, the maximum total number of votes an option could get was

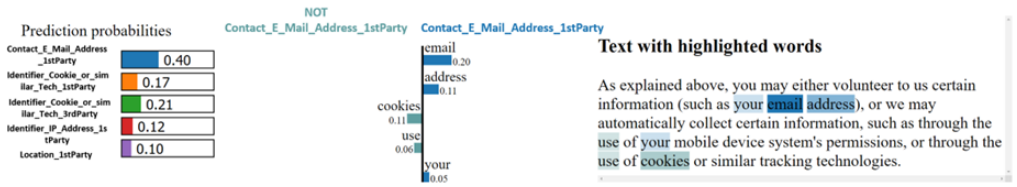


(A) Example explanation 1

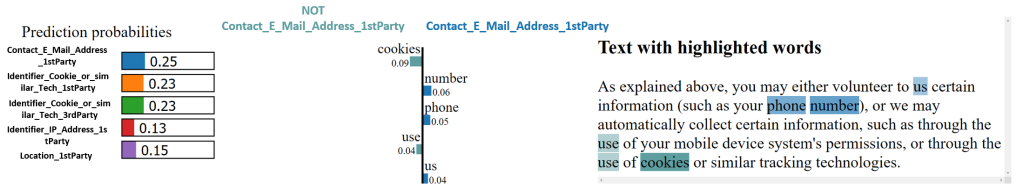


(B) Counterfactual explanation 1

FIGURE 5.3: Example and counterfactual explanation 1



(A) Example explanation 2

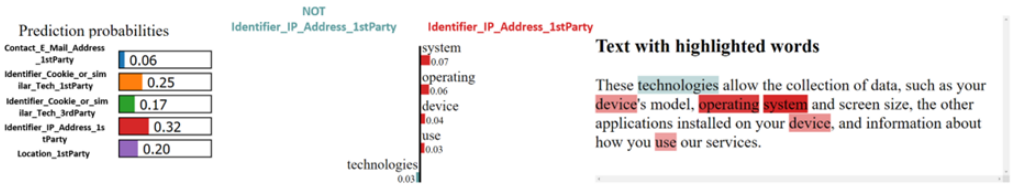


(B) Counterfactual explanation 2

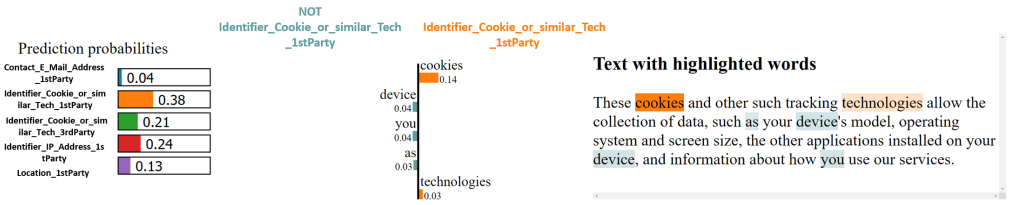
FIGURE 5.4: Example and counterfactual explanation 2

$31 \times 3 = 93$. I totalled up the votes for each option for each part and each question in Figure 5.7 and 5.8.

Overall, respondents found no difference in explainability between logistic regression and SVC (Figure 5.7), while it was more contentious when comparing word representations, with a third split across the three options (Figure 5.8). Remember that the model performance for each pair



(A) Example explanation 3



(B) Counterfactual explanation 3

FIGURE 5.5: Example and counterfactual explanation 3

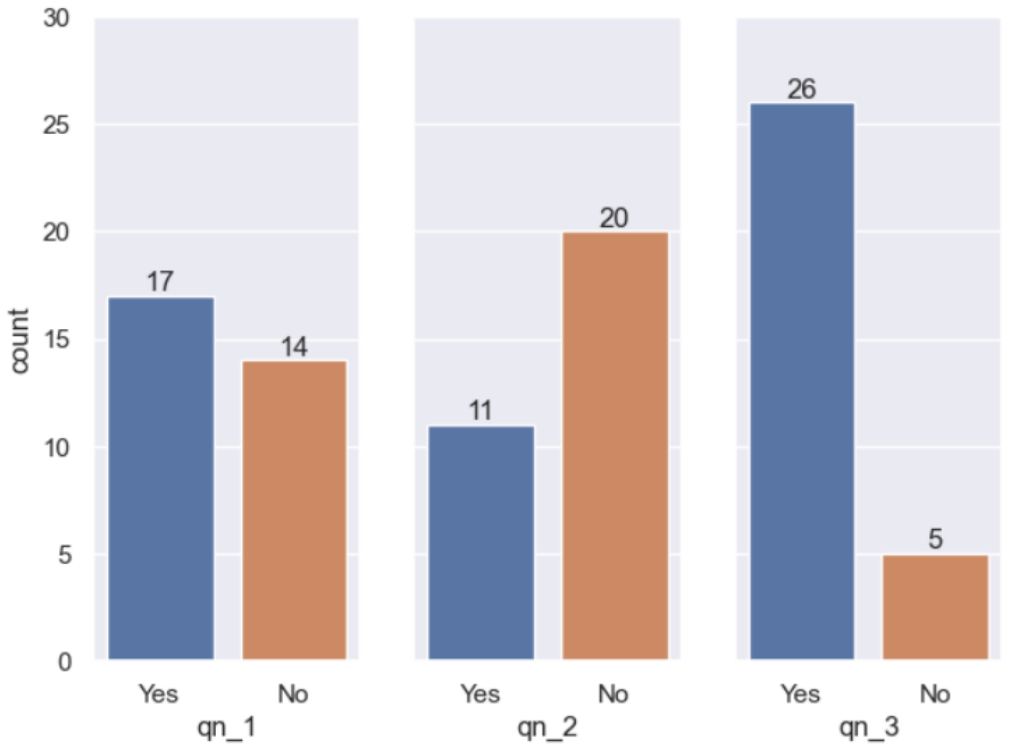


FIGURE 5.6: Votes for predicting whether counterfactual would be classified as `Identifier_Cookie_or_Similar_Tech_1st_Party`

actually did not differ significantly, as mentioned in Chapter 4. In terms of comparing respondents' votes with the "ground truth" metric of model

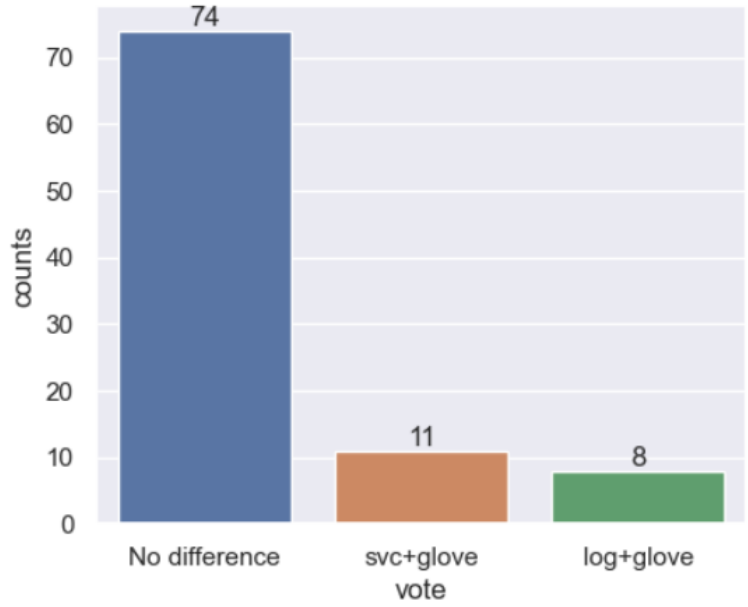
performance, we would expect most respondents to note that there are no differences between the explainabilty of the model. This was seen when comparing logistic regression vs SVC but not seen in the case of comparing Tf-IDF and GloVe. Votes for word representations had a higher spread of votes across the three options. While there is no majority consensus, the fact that the votes were almost equally split instead of being heavily weighted in favour of one category shows that respondents were more undecided in the explainability of the models. One inference of such results is that while respondents collectively did not give a definitive answer as to the more explainable word representation, there are still possible differences in explainability even when comparing models of relatively same performance, as picked up by the differences in distribution of the respondents' votes.

5.1.5 General discussion of results

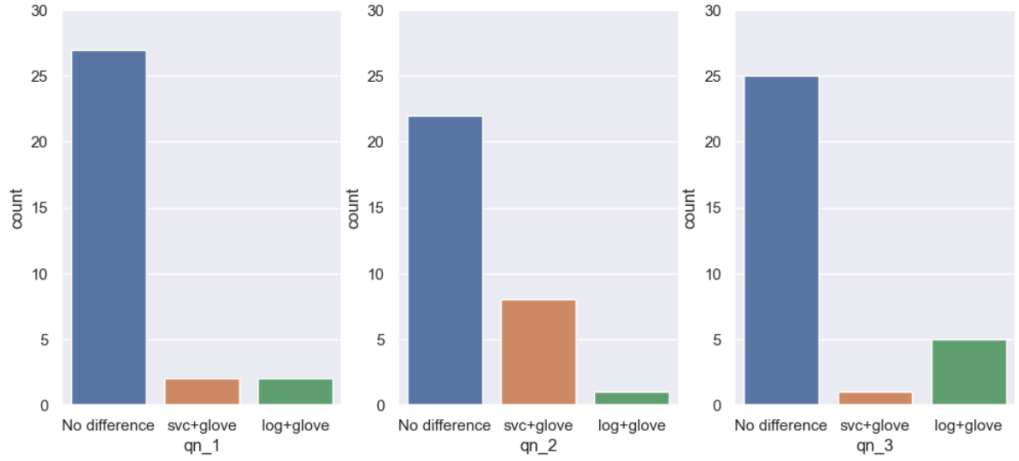
Here are some themes that arise from the preceding discussion:

Explainability is context sensitive.

Whether the self-reported explainability metrics significantly decreased or increased after viewing the explanations depended on the type of metric and the context. Respondents likely placed different emphasis on different values depending on the contexts, which resulted in differing evaluations of the the same explanation. While this finding is uncontroversial, it confirms current XAI research, and also demonstrates the complication of developing "more explainable" models when evaluating such



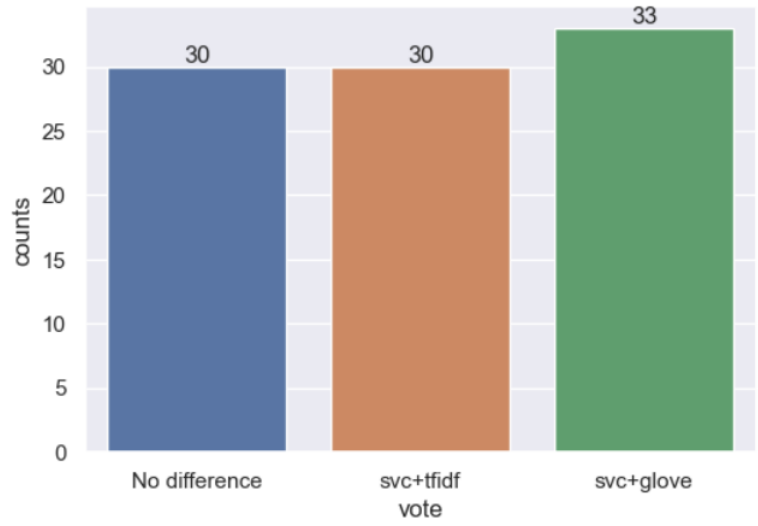
(A) Total votes across 3 questions



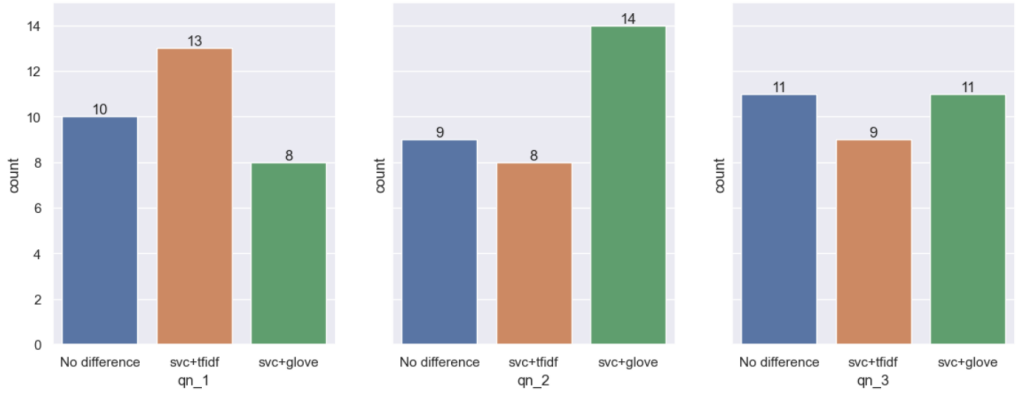
(B) Votes per question

FIGURE 5.7: Testing for which model was more explainable: Respondents' votes to whether SVC + GloVe or Logistic regression + GloVe were more explaninable

explainability is context dependent, compared to typical metrics of performance such as precision / recall / F1. Further, perceptions of explainability can also be influenced by pre-existing expectations that users of models have. If users have the false conception that anything produced by a computer is "objective" and "trustworthy" without being aware that



(A) Total votes across 3 questions



(B) Votes per question

FIGURE 5.8: Testing for which word representation was more explainable: Respondents' votes to whether SVC + TfIDF or SVC + GloVe were more explaninable

machine learning is inherently fallible, they are likely to report a reduction in explainability after viewing explanations that expose the fallibility of these models.

Global and local explainability should be evaluated in tandem.

Global explanability can decrease even though local explanability may

not be an issue. This is one of the limitations of local explainability techniques, as it allows users to understand the model through specific examples. Hence, explainability is highly dependent on the types and number of examples shown to users. If the examples chosen are insufficient or demonstrates an inuitively inconsistent pattern, users may be more confused about the model as a whole which affects global explainability. So it is important to consider global explanability by using different XAI techniques beyond just local explanations. Nevertheless, not all use cases require global explainability, such as perhaps the context of the consumer who may only be concerned with his own instance.

Models with similar performance metrics can have different perceptions of explainability.

Different models and text representations can have different perceptions of explanability, such as the instance when Tf-IDF is compared against GloVe embeddings. While this finding might not surprising given that Tf-IDF is a purely count based metric that does not take into account word context unlike GloVe, such "explanability characteristics" of models can be taken into account when explanability is important for a particular use case beyond simply evaluating based on traditional performance metrics. Nevertheless, it is also important to note that the results obtained here are restricted to the particular dataset and implementation used, and I am not asserting that a particular model or text representation has an "intrinsic explainability quotient".

6 Conclusion

6.1 Findings

We can now answer the research questions that were posed in Section **??**:

1. Which classifiers perform the best on a data privacy dataset in terms of traditional performance metrics?

   Overall, SVC + Tf-IDF performed the best across the 5 data practices according to the weighted F1. This corresponds with the findings of Zimmeck et al. Particularly when classifying `Identifier Cookie 1st Party`, both SVC + Tf-IDF and logistic regression + Tf-IDF performed similarly with the same F1 score. It is surprising that GloVe embeddings performed worse compared to Tf-IDF, given that GloVe accounts for context as compared to Tf-IDF which is purely count-based.

2. Which classifiers are the most explainable to users that include laypersons and users with domain knowledge in data science and the law?

   When SVC was compared to logistic regression, respondents by a large majority found no difference in the explainability. When Tf-IDF was compared to GloVe, respondents were undecided with an almost equal one-third split across the the three options (i.e. no

difference, Tf-IDF & GloVe). So just by looking at these votes, respondents did not find any difference in explainability when comparing SVC to logistic regression, but were undecided whether Tf-IDF or GloVe was more explainable. Comparing the respondents' votes with performance metrics, there seems to be a slight (but non-statistical) relationship between explainability and classifier performance. Tf-IDF consistently outperformed GloVe regardless of the model used, which is reflected in how respondents were equally split between the 3 options. Compared to logistic regression and SVC where latter did not consistently outperform the former, majority of respondents similarly found no difference between the explanations.

3. How would users rate the explainability of a selected XAI technique?

   There was a negative trend of respondents' self-reported scores of interpretability and understanding for LIME. Hence, one possible interpretation is that the individual explanations got less explainable for each question. Another interpretation is that global explainability decreased because respondents were exposed to contradictory information about the model with each new explanation they viewed, rather than local explainability being affected. In terms of predicting counterfactuals, it was not possible to perceive any consistent trend given that the classifiers' predicted probabilities for some classes were almost equal for some of the counterfactual sentences.

4. What are the differences in explainability if users were asked to consider the predictions of these classifiers from the perspective of a consumer, an organisation or the PDPC (as the regulatory authority for data privacy) after viewing the classifiers' explanations?

   Effectiveness and trust significantly decreased for context 2 (PDPC), and fairness for context 3 (consumer). Only effectiveness significantly increased for context 1 (app developer). This could suggest that respondents believe that as a regulatory body, PDPC should retain more discretion in making decisions instead of relying on AI, even though using these models could increase automation. Respondents believe that the risks of making a wrong decision outweighs the possible benefits to efficiency. Having realised the fallibility of the model, respondents have reduced trust as well. Respondents as consumers also seem to be only concerned about their own benefits that result from wrong / correct predictions of the model. Hence, it is not "fair" that they would not be compensated for violations to their data privacy just because the model can make wrong predictions. Respondents as app developers, unlike PDPC, believe that wrong predictions are an acceptable risk when analysing privacy policies can be automated without needing to pay for external legal advice.

   Further, explainability metrics overall significantly decreased for context 2 and 3, which could suggest that respondents felt the risks of making a wrong prediction because of the model's fallibility outweighed any efficiency, or cost gained from automation from the

perspective of PDPC or a consumer. Trust also significantly decreased overall for all contexts. This could suggest that the model's fallibility influenced respondents' trust the most across the contexts, which points to a possible mismatch of expectations of what respondents thought the model could do and what was actually the case.

6.2 Limitations and future work

There are a few limitations:

1. *Limited number of respondents ($n = 31$), diversity of expertise and age*: The intent was to conduct a more in-depth cross sectional study of the self-reported scores of the respondents according to their area and depth of expertise (measured by their major or current occupation) and their beliefs relating to AI and data privacy. This analysis could be used to support certain inferences made earlier. For example, the inference was made that explainability metrics significantly decreased for the contexts of PDPC and consumer but not for the app developer because respondents believed that the app developer should have enough programming expertise to understand the classifier and does not need to rely only on the visualisations provided. To support this claim, a significance test could be used on the scores reported by respondents with data science background to check whether there was a significant decrease, and compared with those respondents without a data science background. However, with 31 respondents, breaking down the survey results

into smaller sub-groups would not reliably show statistical trends[1]. Hence, future work could be conducted on a much larger sample size to run these more in-depth analysis.

2. *Analysis is limited to the context of the APP-350 corpus, classifiers used, and LIME:* As mentioned above, assessing XAI is highly context dependent, and therefore the foregoing analysis should also be seen in light of these particular components. For example, a finding was that respondents found no difference in explainability between SVC and logistic regression. However, this finding is restricted to this particular dataset, as these models and LIME would probably process another sentence from another dataset differently. Future work to evaluate XAI within the context of data privacy could include using another data privacy dataset, or other LIME implementations with different styles of visualisation.

3. *Inherent drawbacks of LIME and human evaluation of XAI:* As LIME is a post-hoc local explainability technique, respondents' understanding of the model was limited to the example visualisations that were presented to them. The sentences and the order of these sentences shown to respondents were intended to show the limitations of the classifier. For example, the first two visualisations showed that "cookie" was the most important feature contributing to the classification, but the next two still classified the sentence as `Identifier Cookie 1st Party` even though "cookie" was not present. This paints a more confusing and inconsistent picture

[1]For example, Wilcoxon Rank Sum Test is recommended to be run on populations > 20.

of the classifier's logic, as compared to four examples consistently showing that "cookie" was the most important feature. In fact, this was controlled for in another study where the authors only chose to show correctly classified sentences in order to reduce the information load on the respondents (Górski and Ramakrishna, 2021). Hence, if more consistent examples were chosen instead, respondents might have reported differently. If post-hoc techniques are used, future work could include an interactive study where respondents can choose to generate their own sentences for classification, which allows them to test whether their understanding of the classifier's logic is consistent.

Further, human evaluations of post-hoc explanations have been criticised for being inherently flawed because of confirmation bias. Post-hoc explanations may have little to no similarities to how the underlying classifier made the prediction as they are an oversimplification of the classifier's logic. Hence, human evaluation may have limited meaning (Rosenfeld, 2021). Instead, the authors propose four objective metrics that quantify the explanation itself and its appropriateness given the XAI goal. Future work could involve less reliance on human evaluation to incorporate more objective metrics, as well as choosing other types of XAI techniques such as global self explaining techniques.

4. *LinearSVC and logistic regression can be similar in their optimisation technique, which may influence explainability.*

6.3 Final comments

(Hacker and Passoth, 2022) - about intersection of law and explainability in different areas of law

(Yalcin et al., 2022) - reason for why efficiency and trust is emphasised in this capstone

(Vilone and Longo, 2021) - about how there are so many different metrics as part of explainability, including efficiency, trust, transparency, understandability etc.

(Kästner et al., 2021) - about the relationship between trust and explainability - can be used to explain survey results

(Waltl and Vogl, 2018)- Possible to assign metrics of explainability to AI models (like a rating)

Bibliography

Alkaissi, Hussam and Samy I McFarlane (2023). "Artificial hallucinations in chatgpt: Implications in scientific writing". In: *Cureus* 15.2.

Ashley, Kevin D. (2017). *Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age*. Cambridge University Press. DOI: 10.1017/9781316761380.

Chesterman, Simon (2021a). "We, The Robots: Regulating Artificial Intelligence and the Limits of the Law". In: Cambridge University Press. Chap. 6. Transparency.

— (2021b). "We, The Robots: Regulating Artificial Intelligence and the Limits of the Law". In: Cambridge University Press. Chap. 3. Opacity.

Danilevsky, Marina et al. (2020). "A Survey of the State of Explainable AI for Natural Language Processing". In: *CoRR* abs/2010.00711. arXiv: 2010.00711. URL: https://arxiv.org/abs/2010.00711.

Doshi-Velez, Finale and Been Kim (2017). "Towards a Rigorous Science of Interpretable Machine Learning". In: *arXiv preprint arXiv:1702.08608*.

Dunning, David (2011). "Chapter five - The Dunning–Kruger Effect: On Being Ignorant of One's Own Ignorance". In: ed. by James M. Olson and Mark P. Zanna. Vol. 44. Advances in Experimental Social Psychology. Academic Press, pp. 247–296. DOI: https://doi.org/10.1016/B978-0-12-385522-0.00005-6. URL: https://www.sciencedirect.com/science/article/pii/B9780123855220000056.

El Zini, Julia and Mariette Awad (2022). "On the Explainability of Natural Language Processing Deep Models". In: *ACM Computing Surveys* 55.103, pp. 1–31. DOI: https://doi.org/10.1145/3529755.

Górski, Łukasz and Shashishekar Ramakrishna (2021). "Explainable Artificial Intelligence, Lawyer's Perspective". In: *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*. ICAIL '21. São Paulo, Brazil: Association for Computing Machinery, 60–68. ISBN: 9781450385268. DOI: 10.1145/3462757.3466145. URL: https://doi.org/10.1145/3462757.3466145.

Greenstein, Stanley (2022). "Preserving the rule of law in the era of artificial intelligence (AI)". In: *Artificial Intelligence and Law* 30.3, pp. 291–323.

Gstrein, Oskar J. and Anne Beaulieu (2022). "How to protect privacy in a datafied society? A presentation of multiple legal and conceptual approaches". In: *Philos Technol* 35.1, p. 3. DOI: https://doi.org/10.1007%2Fs13347-022-00497-4.

Hacker, Philipp and Jan-Hendrik Passoth (2022). "Varieties of AI Explanations Under the Law. From the GDPR to the AIA, and Beyond". In: *xxAI-Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*. Springer, pp. 343–373.

Hacker, Philipp et al. (2020). "Explainable AI under contract and tort law: legal incentives and technical challenges". In: *Artificial Intelligence and Law* 28, pp. 415–439.

Hsu, Bi-Min (2020). "Comparison of supervised classification models on textual data". In: *Mathematics* 8.5, p. 851.

IBM (2022). *What is natural language processing?* Last Accessed: 2023-02-11. URL: https://www.ibm.com/sg-en/topics/natural-language-processing.

Kästner, Lena et al. (2021). "On the relation of trust and explainability: Why to engineer for trustworthiness". In: *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)*. IEEE, pp. 169–175.

Katz, Daniel Martin et al. (2023a). "GPT-4 Passes the Bar Exam". In: *Available at SSRN 4389233*.

Katz, Daniel Martin et al. (2023b). "Natural Language Processing in the Legal Domain". In: Available at SSRN. DOI: https://dx.doi.org/10.2139/ssrn.4336224.

Linardatos, Pantelis, Vasilis Papastefanopoulos, and Sotiris Kotsiantis (2020). "Explainable AI: A Review of Machine Learning Interpretability Methods". In: *Entropy (Basel)* 23.1, p. 18. DOI: https://dx.doi.org/10.3390/e23010018.

Liu, Zhiyuan et al. (2020). "Word representation". In: *Representation Learning for Natural Language Processing*, pp. 13–41.

Mantelero, Alessandro (2014). "The future of consumer data protection in the EU Re-thinking the "notice and consent" paradigm in the new era of predictive analytics". In: *Computer Law & Security Review* 30.6, pp. 643–660.

OpenAI (2022). *ChatGPT: Optimizing Language Models for Dialogue*. Last Accessed: 2023-02-11. URL: https://openai.com/blog/chatgpt/.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning (2014). "GloVe: Global Vectors for Word Representation". In: *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543. URL: http://www.aclweb.org/anthology/D14-1162.

Personal Data Protection Commission (Jan. 2020). *Model Artificial Intelligence Governance Framework Second Edition*. Tech. rep. Last Accessed: 2023-2-20. Personal Data Protection Commission. URL: https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/sgmodelaigovframework2.pdf.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016a). *LIME*. URL: https://github.com/marcotcr/lime.

— (2016b). *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. DOI: 10.48550/ARXIV.1602.04938. URL: https://arxiv.org/abs/1602.04938.

Rosenfeld, Avi (2021). "Better Metrics for Evaluating Explainable Artificial Intelligence". In: *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*. AAMAS '21. Virtual Event, United Kingdom: International Foundation for Autonomous Agents and Multiagent Systems, 45–50. ISBN: 9781450383073.

SciPy (2023). *scipy.stats.wilcoxon*. Last Accessed: 2023-3-1. URL: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.wilcoxon.html.

Vale, Daniel, Ali El-Sharif, and Muhammed Ali (2022). "Explainable artificial intelligence (XAI) post-hoc explainability methods: Risks and limitations in non-discrimination law". In: *AI and Ethics*, pp. 1–12.

Vilone, Giulia and Luca Longo (2021). "Notions of explainability and evaluation approaches for explainable artificial intelligence". In: *Information Fusion* 76, pp. 89–106.

Vylomova, Ekaterina et al. (2015). *Take and Took, Gaggle and Goose, Book and Read: Evaluating the Utility of Vector Differences for Lexical Relation Learning*. DOI: 10.48550/ARXIV.1509.01692. URL: https://arxiv.org/abs/1509.01692.

Wagner, Isabel (2022). "Privacy Policies Across the Ages: Content and Readability of Privacy Policies 1996–2021". In: *arXiv preprint arXiv:2201.08739*.

Waltl, Bernhard and Roland Vogl (2018). "Explainable artificial intelligence the new frontier in legal informatics". In: *Jusletter IT* 4, pp. 1–10.

Yalcin, Gizem et al. (2022). "Perceptions of justice by algorithms". In: *Artificial Intelligence and Law*, pp. 1–24.

Zednik, Carlos (2021). "Solving the black box problem: A normative framework for explainable artificial intelligence". In: *Philosophy & technology* 34.2, pp. 265–288.

Zimmeck, Sebastian et al. (2019). "MAPS: Scaling Privacy Compliance Analysis to a Million Apps". In: *Proceedings on Privacy Enhancing Technologies* 2019.3, pp. 66–86.

A Survey Questions

Survey: Using Explainable AI (XAI) Techniques on a Data Privacy dataset

Thanks for taking this survey! As part of my capstone project, I aim to assess the effectiveness of several machine learning model explanations and your sentiments & opinions about decisions made by artificial intelligence (AI).

Please note that there are no "correct" answers to the questions that follow, simply choose whichever option that you agree with the most.

This survey takes approximately 15 minutes to complete. I highly recommend you complete it on a desktop / laptop within one sitting.

* Required

1. If you are a NUS / Yale-NUS student, are you 18 years old and above? *

OR

If you are not a NUS student, are you above 21+ years old?

These age restrictions are due to Yale-NUS ethics guidelines.

Mark only one oval.

- ◯ Yes
- ◯ No

Participant Information Sheet

The following information contains details about this capstone. After reading the details, click "Yes" at the bottom of this section if you wish to participate in this research.

1. Research Project Title
Applying Explainable AI techniques onto the APP-350 Corpus

2. Who is this researcher?
I am a fifth-year Double Degree in Law and Liberal Arts student at Yale-NUS College, supervised by Professor Michael Choi and Professor Simon Chesterman.

3. What is the purpose of this research?
You are invited to participate in a research study. This sheet provides you with information about the research study. The researcher identified above will describe this research to you and answer all of your questions. Read the information below before deciding whether or not to take part. My capstone aims to train and assess explainable AI models that are able to classify legal text. Part of the explanability of these models include visualisations of how the model classifies text. I aim to survey law and non-law students on how interpretable these visualisations are. I will ask you whether you find these visualisations understandable, and whether you think these are reasonable explanations of why the AI model made a certain prediction. I will also ask questions that survey how much you trust the AI model as a result of these explanations.

4. Who can participate in the research? What is the expected duration of my participation? What is the duration of this research?
18+ (NUS / Yale-NUS students), and 21+ for the general public. The time demand is about 10 – 15 minutes, and the overall duration of study is from December 2022 – May 2023.

5. What will happen to me if I take part?
You will answer questions on a Google form. I aim to survey law and non-law students on how interpretable these visualisations are. I will ask you whether you find these visualisations understandable, and whether you think these are reasonable explanations of why the AI model made a certain prediction. I will also ask questions that survey how much you trust the AI model as a result of these explanations.

6. If I agree to take part, what happens to the data I provide?
Collected data will be kept confidential and will only be used for the stated research, which will take place between August 2022 and May 2023. Unidentified quotes may be used in presentations or the capstone report, which will not contain your name. All data (not including personal identifiers, such as names and contact information) collected will be kept in accordance with the University's Research Data Management Policy. Research data used in any publication will be kept for a minimum of 10 years before being discarded.

7. How will my privacy and the confidentiality of my research records be protected?
No personal identifiable information will be recorded as this is an online survey.

8. What are the possible discomforts and risks for participants?
No discomfort or risk is expected with this research.

9. What is the compensation for any injury?
If you follow the directions of the researcher in charge of this research study and you are injured, the NUS will pay the medical expenses for the treatment of that injury. By giving your consent, you will not waive any of your legal rights or release the parties involved in this study from liability for negligence.

10. Will there be reimbursement for participation?
No reimbursement will be given.

11. What are the possible benefits to me and to others?
There is no direct benefit to you by participating in this research study.

12. Can I refuse to participate in this research?
Yes, you can. Your decision to participate in this research study is voluntary and completely up to you. You can also withdraw from the research at any time without giving any reasons, by informing the researcher or supervisor, after which all of your data collected will be discarded.

13. Whom should I call if I have any questions or problems?
Name of Researcher: Tristan Koh
Contact number: 96219123
Email: tristan.koh@u.yale-nus.edu.sg
Supervisor's Name: Michael Choi
Supervisor's email: tristan.koh@u.yale-nus.edu.sg
Secretariat, College Ethics Review Committee: researchethics@yale-nus.edu.sg

Thank you for reading this information sheet and for considering taking part in this research.

2. I have read about the purpose of this research study, agree to participate, and understand that I can withdraw at any time. *

Mark only one oval.

- ◯ Yes
- ◯ No

Part 1

I would like to capture some demographic information as well as your beliefs and views of artificial intelligence (AI).

3. What is your major / prospective major? *

   If you are in a double degree program, choose whichever degree you have more experience in.

   If you have graduated, choose the subject area which best corresponds to your expertise in your current work.

   Mark only one oval.

   - ◯ Law
   - ◯ MCS / Computer Science / Data Science / Statistics
   - ◯ Accountancy
   - ◯ Anthropology
   - ◯ Arts & Humanities
   - ◯ Business
   - ◯ Economics
   - ◯ Engineering
   - ◯ Environmental Studies
   - ◯ Global Affairs / Political Science
   - ◯ History
   - ◯ Literature
   - ◯ Life Sciences
   - ◯ Medicine
   - ◯ Physical Sciences
   - ◯ Philosophy
   - ◯ Psychology
   - ◯ PPE
   - ◯ PFM
   - ◯ Urban Studies
   - ◯ Other: _____

4. What is your age? *

   Mark only one oval.

   - ◯ 18 - 25
   - ◯ 26 - 35
   - ◯ 36 - 45
   - ◯ 46 - 55
   - ◯ 56+

5. Do you have any experience with AI / data science / programming? *

Mark only one oval.

None at all

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Very experienced

6. Do you have any experience with regarding data privacy / law? *

Mark only one oval.

None at all

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Very experienced

7. Are you concerned about your data privacy? *

Mark only one oval.

Not concerned at all

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Very concerned

8. How would you rate your capability in protecting your online data? *

Mark only one oval.

Not capable at all

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Very capable

9. Do you think decisions that are made by AI can be useful to society? *

In answering this question and later questions, please use whichever meaning of "useful" you agree with the most.

Mark only one oval.

Not useful at all

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Very useful

10. Do you think decisions made by AI are fair? *

In answering this question and later questions, please use whichever meaning of "fair" you agree with the most.

Mark only one oval.

Very unfair

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Perfectly fair

11. Do you think decisions made by AI can be a risk to society? *

In answering this question and later questions, please use whichever meaning of "risk" you agree with the most.

Mark only one oval.

	No risk to society
1	
2	
3	
4	
5	
	High risk to society

Part 2

To answer the rest of the questions, I would like you to read the text below to understand the context of the capstone. Don't worry if you are not able to understand everything, as this survey is meant to assess laypeople's understanding of AI.

What are data privacy policies?

- Every app that you download contains a data privacy policy that states how the app will collect and use your personal data.
- By using the app, you consent to the data privacy policy.
- Due to the Personal Data Protection Act (PDPA), app developers have to notify you and ask for your consent if they collect and use your personal data. If they fail to do so, they would be in violation of the PDPA.

How does AI work?

- The machine learning model ("model") in this capstone is used to predict whether sentences from an app's data privacy policy fall into a particular data practice. A data practice is what the app does with the user's data.
- For example, consider the sentence: "In connection with these advertising services, we may use cookies, web beacons, and similar technologies to collect behavioral information about how you use our site or other websites in order to perform tracking and marketing analytics or serve advertisements that are more likely to be of interest to you."
- Cookies are files created by websites you visit. They make your online experience easier by saving browsing information.
- This sentence is classified as "*Identifier_Cookie_or_similar_Tech_1stParty*" because the sentence states that the app uses cookies (or other tracking technologies) to track the user's activities. This sentence is also classified as "1stParty" as the data is only collected by the app, and not shared with other organisations.

- However, as the models make predictions by generalising from examples given to the model, not all the model's predictions will be correct. The model could classify the sentence wrongly as another data privacy practice, or it could also classify it as not containing any data privacy practice.
- For example, consider the sentence: "These technologies also enable us to provide features such as storage of items in your cart between visits and Short Message Service (SMS)/text messages you have chosen to receive."

- Even though the sentence does not contain the word "cookies", similar tracking technology are still being used because the app is able to track items in the user's cart in between visits. Therefore, tracking technology is still being used even though the word "cookies" is not specifically stated.
- Therefore, if the model relies heavily on "cookies" as a key word to correctly classify the sentence, the model's prediction would likely be wrong as the sentence does not contain "cookies".

12. Please select "strongly agree" to show that you are paying attention to this question. *

Mark only one oval.

- ◯ Strongly Agree
- ◯ Agree
- ◯ Neutral
- ◯ Disagree
- ◯ Strongly Disagree

In this section, I will describe three different contexts with similar facts that relate to the use of the abovementioned model in analysing data privacy policies.

Each context corresponds with the perspective of an app developer, a member of the Personal Data Protection Commission (PDPC), and an user of the app.

I would then ask you questions to capture how your opinions on the use of AI in decision making would differ based on these three different perspectives.

Context 1: Imagine that you are an app developer. You are developing an app that uses cookies to track user activity online. To comply with the PDPA, you know that you need to include a sentence in your app's data privacy policy that notifies and asks for users' consent to use cookies.

Since you have no knowledge of the PDPA, you use the abovementioned model to analyse a pre-drafted data privacy policy that you found online. The model informs you that there is a sentence which states that cookies are being used.

You are deciding whether to rely entirely on the model's prediction, or pay costly legal fees to confirm with your friend who is a lawyer.

If the pre-drafted data privacy policy actually does not state that cookies are being used but your app uses cookies, you could face a fine of up to $10,000 in breach of the PDPA as you would have failed to notify your users.

How far do you, as the app developer:

13. Think that using the model is an effective method of identifying violations of the PDPA? *

In answering this question and later questions, please use whichever meaning of "effective" you believe is the most appropriate for the context.

Mark only one oval.

Not effective at all

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Very effective

14. Think that using the model is a fair method of identifying violations of the PDPA? *

In answering this question and later questions, please use whichever meaning of "fair" you believe is the most appropriate for the context.

Mark only one oval.

Very unfair

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Very fair

15. Think that using the model is a method that could be a risk to society? *

In answering this question and later questions, please use whichever meaning of "risk to society" you believe is the most appropriate for the context.

Mark only one oval.

Very risky

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Not risky

16. Trust the prediction made by the model? *

In answering this question and later questions, please use whichever meaning of "trust" you believe is the most appropriate for the context.

Mark only one oval.

Do not trust at all

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Trust very much

Context 2: Imagine that you are a committee member part of the Personal Data Protection Commission (PDPC). A user of an app has informed you that an app is using cookies but has not notified its users.

Your team checks the code of the app and confirms that the app is indeed using cookies. Your team uses the abovementioned model and the model informs you that the data privacy policy does not contain any sentence that notifies its users that it uses cookies.

To increase the efficiency of the PDPC, your team is considering whether to adopt the abovementioned model to automate the analysis of data privacy policies. If this new method of analysis is adopted, the PDPC would rely entirely on the model's predictions to confirm whether app developers have breached the PDPA. The app developers would face a fine of up to $10,000 if they are found to have breached the PDPA.

How far would you, as a committee member of the PDPC:

17. Think that using the model is an effective method of identifying violations of the PDPA? *

Mark only one oval.

Not effective at all

1

2

3

4

5

Very effective

18. Think that using the model is a fair method of identifying violations of the PDPA? *

Mark only one oval.

Very unfair

1

2

3

4

5

Very fair

19. Think that using the model is a method that could be a risk to society? *

Mark only one oval.

Very risky

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Not risky

20. Trust the prediction made by the model? *

Mark only one oval.

Do not trust at all

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Trust very much

Context 3: Imagine that you are a user of an app. You read in a forum where other users allege that the app uses cookies. You decide to analyse the data privacy policy of the app using the abovementioned model and the model informs you that the data privacy policy does not contain any sentence that notifies its users that it uses cookies.

You are deciding whether to submit this prediction as the only supporting piece of evidence to the PDPC to claim that the app has used cookies without notifying you.

If the PDPC decides that the developer has indeed violated the PDPA, you could claim compensation from the app developer of up to $10,000.

How far would you, as a user of the app:

21. Think that using the model is an effective method of identifying violations of the PDPA? *

Mark only one oval.

Not effective at all

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Very effective

22. Think that using the model is a fair method of identifying violations of the PDPA? *

Mark only one oval.

Very unfair

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Very fair

23. Think that using the model is a method that could be a risk to society? *

Mark only one oval.

Very risky

1 ◯
2 ◯
3 ◯
4 ◯
5 ◯

Not risky

24. Trust the prediction made by the model? *

Mark only one oval.

Do not trust at all

1 ◯
2 ◯
3 ◯
4 ◯
5 ◯

Trust very much

Part 3

In the next 3 sections, I would like to capture your opinions about the understandability of the model that I have been using in this capstone. Please carry on reading to learn more about the specifics of the model.

As above, don't worry if you are not able to understand everything, as this survey is meant to assess laypeople's understanding of AI.

What is this model used to predict?
This section involves predictions made by a model trained on sentences with the data practice "Identifier_Cookie_or_similar_Tech_1stParty".

The definition of this data practice is that:

- The app uses cookies (or other tracking technologies) to track the user's activities.
- The data is only collected by the app itself, and not shared with other organisations.

The following sentences are some illustrative sentences that were predicted correctly by this model (i.e. the predicted practice was correctly predicted to be "Identifier_Cookie_or_similar_Tech_1stParty"):

- We automatically receive and track certain information about your computer or mobile device when you visit our sites or apps, including through the use of cookies.
- However, if you block or erase cookies, we may not be able to restore any preferences or customisation settings you have previously specified, and our ability to personalise your online experience would be limited.
- Other technologies, such as Silverlight storage, may be cleared from within the application.

However, the following sentences are also illustrative examples of sentences which were classified wrongly by the model (i.e. the practice was not predicted to be "Identifier_Cookie_or_similar_Tech_1stParty"):

- In connection with these advertising services, we or our Advertising Service Providers, like Google Analytics may use cookies, web beacons, and similar technologies to collect behavioral information about how you use our site or other websites in order to perform tracking and marketing analytics or serve advertisements that are more likely to be of interest to you.
- Shared Information also includes information about you (including Location Data and Log Data) that others who are using our services share about you.
- As explained above, you may either volunteer to us certain information (such as your email address), or we may automatically collect certain information, such as through the use of your mobile device system's permissions, or through the use of cookies or similar tracking technologies.

Out of 425 sentences, the model:

1. correctly classified 311 sentences; and
2. wrongly classified 114 sentences.
3. The accuracy of the model is 73%.

Note that the model was also trained to identify four other data practices. Hence, you would see other practices stated in some visualisations when the model predicted the wrong practice.

25. Do you understand why the model made the prediction? *

Mark only one oval.

Don't understand at all

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Fully understand

26. Why do you think the model made this prediction? *

A short answer would suffice, there is no "correct" answer.

Instructions for this section

In this section, you will be asked to assess the effectiveness of visualisations that explain how the abovementioned model makes predictions.

You will also be told whether the model predicted the practice correctly / wrongly, and be given the predicted practice by the model and actual practice, such as:

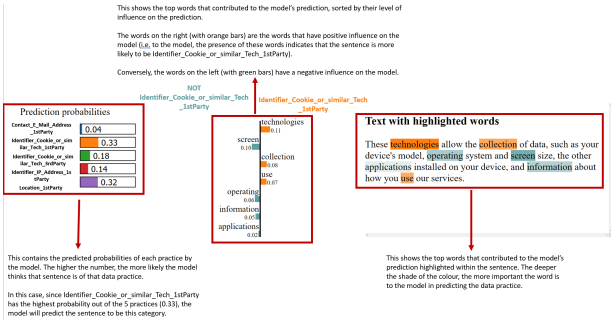
The model predicted the practice wrongly.

Predicted practice: Identifier_Cookie_or_Similar_Tech_1stParty

Actual practice: Identifier_IP_Address

Please note that there are no "correct" answers, simply choose the option which you think you agree with the most.

The following image explains how to interpret a visualisation.



Visualisation 3.1

The model predicted the practice correctly.

Predicted practice: Identifier_Cookie_or_Similar_Tech_1stParty

Actual practice: Identifier_Cookie_or_Similar_Tech_1stParty

27. Did you find the visualisation easy to interpret? *

Mark only one oval.

	Very difficult
1	◯
2	◯
3	◯
4	◯
5	◯

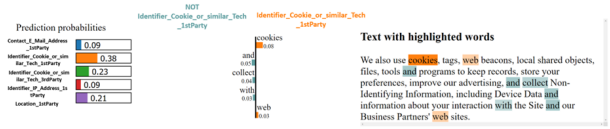
Very easy

Visualisation 3.2

The model predicted the practice correctly.

Predicted practice: Identifier_Cookie_or_Similar_Tech_1stParty

Actual practice: Identifier_Cookie_or_Similar_Tech_1stParty



28. Do you understand why the model made the prediction? *

Mark only one oval.

Don't understand at all

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Fully understand

29. Why do you think the model made this prediction? *

A short answer would suffice, there is no "correct" answer.

30. Did you find the visualisation easy to interpret? *

Mark only one oval.

Very difficult

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Very easy

31. Based on your current understanding, do you think that the sentence below would be predicted to be "Identifier_Cookie_or_Similar_Tech_1stParty"? *

"We also use **tracking technologies** to keep records, store your preferences, improve our advertising, and collect Non-Identifying Information, including Device Data and information about your interaction with the Site and our Business Partners' web sites."

The difference in the sentence in the visualisation above and the provided sentence is highlighted in **bold**.

Mark only one oval.

- ◯ Yes
- ◯ No

Visualisation 3.3

The model predicted the practice wrongly.

Predicted practice: Contact_E_Mail_Address_1stParty

Actual practice: Identifier_Cookie_or_Similar_Tech_1stParty

Note: Contact_E_Mail_Address_1stParty is a data practice where the app collects the email address of the app user.

32. Do you understand why the model made the prediction? *

Mark only one oval.

Don't understand at all

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Fully understand

33. Why do you think the model made this prediction? *

A short answer would suffice, there is no "correct" answer.

34. Did you find the visualisation easy to interpret? *

Mark only one oval.

Very difficult

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Very easy

35. Based on your current understanding, do you think the sentence below would be predicted to be in "Identifier_Cookie_or_Similar_Tech_1stParty?" *

"As explained above, you may either volunteer to us certain information (such as your **phone number**), or we may automatically collect certain information, such as through the use of your mobile device system's permissions, or through the use of cookies or similar tracking technologies."

The difference in the sentence in the visualisation and the provided sentence is highlighted in **bold**.

Mark only one oval.

- ◯ Yes
- ◯ No

Visualisation 3.4

The model predicted the practice wrongly.

Predicted practice: Identifier_IP_Address_1stParty

Actual practice: Identifier_Cookie_or_Similar_Tech_1stParty

Note: Identifier_IP_Address_1stParty is a data practice where the app collects the user's IP address.

36. Do you understand why the model made the prediction? *

Mark only one oval.

Don't understand at all

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Fully understand

37. Why do you think the model made this prediction? *

A short answer would suffice, there is no "correct" answer.

38. Did you find the visualisation easy to interpret? *

Mark only one oval.

Very difficult

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Very easy

39. Based on your current understanding, do you think the sentence below would be predicted to be in "Identifier_Cookie_or_Similar_Tech_1stParty?" *

"These **cookies and other such tracking technologies** allow the collection of data, such as your device's model, operating system and screen size, the other applications installed on your device, and information about how you use our services."

The difference in the sentence in the visualisation and the provided sentence is highlighted in **bold**.

Mark only one oval.

- ◯ Yes
- ◯ No

Part 4

In this section, you will be given pairs of different visualisations of the predictions of the same sentence. For each visualisation, I will state whether the model predicted the practice correctly / wrongly, and also state the predicted practice by the model and the actual practice.

Choose the visualisation that is more understandable to you on first glance. If both visualisations seem the same to you, then select "no difference".

Once again, there is no "correct" option. Simply choose whichever option that you agree with the most.

Visualisation 4.1(i)

The model predicted the practice correctly.

Predicted practice: Identifier_Cookie_or_similar_Tech_1stParty

Actual practice: Identifier_Cookie_or_similar_Tech_1stParty

Visualisation 4.1(ii)

The model predicted the practice correctly.

Predicted practice: Identifier_Cookie_or_similar_Tech_1stParty

Actual practice: Identifier_Cookie_or_similar_Tech_1stParty

40. W * hich example annation did you find easier to interpret?

Mark only one oval.

- 4.1(i)
- 4.1(ii)

N
o
d
i
f
f
e
r
e
n
c
e

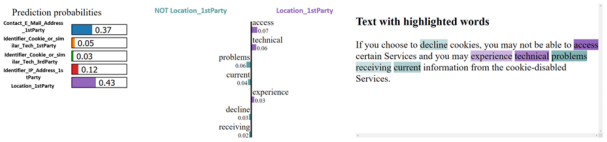
Visualisation 4.2(i)

The model predicted the practice wrongly.

Predicted practice: Location_1stParty

Actual practice: Identifier_Cookie_or_similar_Tech_1stParty

Note: Location_1stParty is a data practice where the app collects the location data of the user.



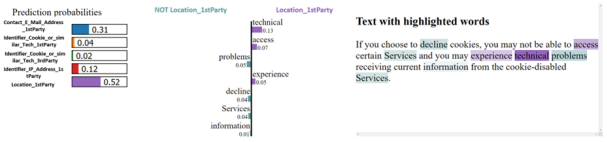
Visualisation 4.2(ii)

The model predicted the practice wrongly.

Predicted practice: Location_1stParty

Actual practice: Identifier_Cookie_or_similar_Tech_1stParty

Note: Location_1stParty is a data practice where the app collects the location data of the user.



41. Which explanation did you find easier to interpret? *

Mark only one oval.

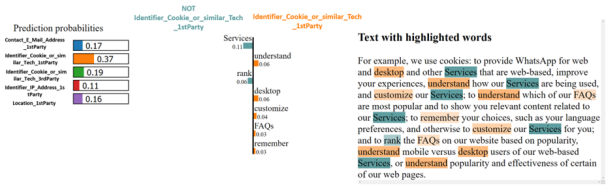
- ◯ 4.2(i)
- ◯ 4.2(ii)
- ◯ No difference

Visualisation 4.3(i)

The model predicted the practice correctly.

Predicted practice: Identifier_Cookie_or_similar_Tech_1stParty

Actual practice: Identifier_Cookie_or_similar_Tech_1stParty

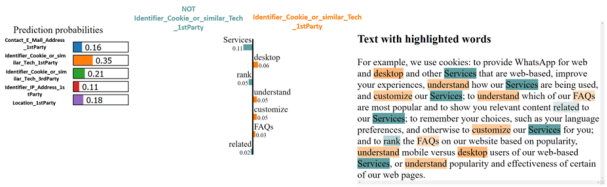


Visualisation 4.3(ii)

The model predicted the practice correctly.

Predicted practice: Identifier_Cookie_or_similar_Tech_1stParty

Actual practice: Identifier_Cookie_or_similar_Tech_1stParty



42. Which explanation did you find easier to interpret? *

Mark only one oval.

- 4.3(i)
- 4.3(ii)
- No difference

Part 5

As with the previous section, you will be given pairs of different visualisations of the predictions of the same sentence. Choose the visualisation that is more understandable to you on first glance.

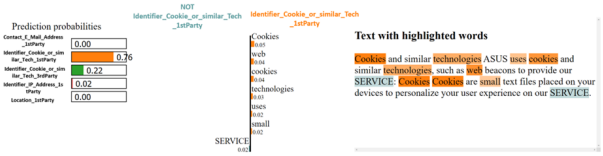
Choose "no difference" if the visualisations seem the same to you.

Visualisation 5.1(i)

The model predicted the practice correctly.

Predicted practice: Identifier_Cookie_or_similar_Tech_1stParty

Actual practice: Identifier_Cookie_or_similar_Tech_1stParty

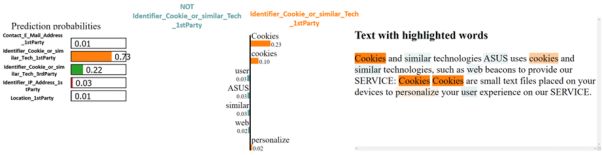


Visualisation 5.1(ii)

The model predicted the practice correctly.

Predicted practice: Identifier_Cookie_or_similar_Tech_1stParty

Actual practice: Identifier_Cookie_or_similar_Tech_1stParty



43. Which explanation did you find easier to interpret? *

Mark only one oval.

- ◯ 5.1(i)
- ◯ 5.1(ii)
- ◯ No difference

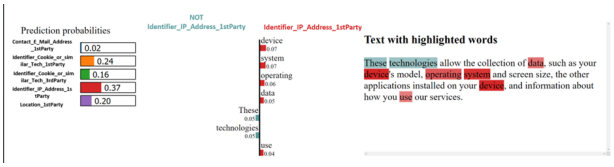
Visualisation 5.2(i)

The model predicted the practice wrongly.

Predicted practice: Identifier_IP_Address_1stParty

Actual practice: Identifier_Cookie_or_similar_Tech_1stParty

Note: Identifier_IP_Address_1stParty is a data practice where the app collects the user's IP address.

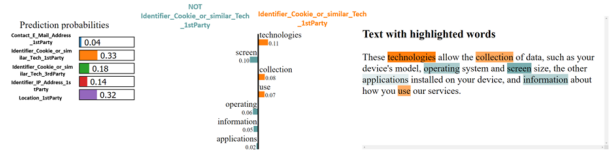


Visualisation 5.2(ii)

The model predicted the practice correctly.

Predicted practice: Identifier_Cookie_or_similar_Tech_1stParty

Actual practice: Identifier_Cookie_or_similar_Tech_1stParty



44. Which explanation did you find easier to interpret? *

Mark only one oval.

- ◯ 5.2(i)
- ◯ 5.2(ii)
- ◯ No difference

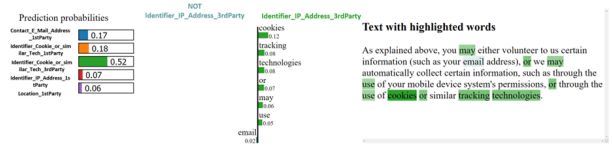
Visualisation 5.3(i)

The model predicted the practice wrongly.

Predicted practice: Identifier_Cookie_or_similar_Tech_3rdParty

Actual practice: Identifier_Cookie_or_similar_Tech_1stParty

Identifier_Cookie_or_similar_Tech_3rdParty is a practice where the app uses cookies or similar tracking technologies and shares it with third party organisations.

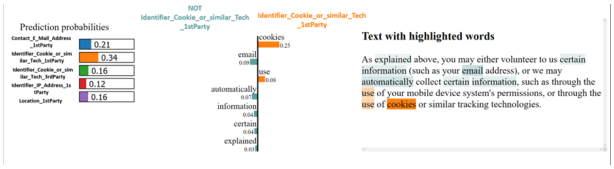


Visualisation 5.3(ii)

The model predicted the practice correctly.

Predicted practice: Identifier_Cookie_or_similar_Tech_1stParty

Actual practice: Identifier_Cookie_or_similar_Tech_1stParty



45. Which explanation did you find easier to interpret? *

Mark only one oval.

- 5.3(i)
- 5.3(ii)
- No difference

Part 6

This section asks you to respond to the same questions in the same three contexts (i.e. app developer, member of the PDPA, app user) as described in Part 3.

Similarly, please consider how your opinions about the use of AI in decision making changes with the different perspectives of these three roles.

Please answer these following questions taking into account your current understanding of the model having viewed the explanations in the previous parts.

Context 1: Imagine that you are an app developer. You are developing an app that uses cookies to track user activity online. To comply with the PDPA, you know that you need to include a sentence in your app's data privacy policy that notifies and asks for users' consent to use cookies.

Since you have no knowledge of the PDPA, you use the abovementioned model to analyse a pre-drafted data privacy policy that you found online. The model informs you that there is a sentence which states that cookies are being used.

You are deciding whether to rely entirely on the model's prediction, or pay costly legal fees to confirm with your friend who is a lawyer.

If the pre-drafted data privacy policy actually does not state that cookies are being used but your app uses cookies, you could face a fine of up to $10,000 in breach of the PDPA as you would have failed to notify your users.

How far do you, as the app developer:

46. Think that using the model is an effective method of identifying violations of the PDPA? *

Mark only one oval.

Not effective at all

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Very effective

47. Think that using the model is a fair method of identifying violations of the PDPA? *

Mark only one oval.

Very unfair

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Perfectly fair

48. Think that using the model is a method that could be a risk to society? *

Mark only one oval.

Very risky

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Not risky

49. Trust the prediction made by the model? *

Mark only one oval.

Do not trust at all

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Trust very much

Context 2: Imagine that you are a committee member part of the Personal Data Protection Commission (PDPC). A user of an app has informed you that an app is using cookies but has not notified its users.

Your team checks the code of the app and confirms that the app is indeed using cookies. Your team uses the abovementioned model and the model informs you that the data privacy policy does not contain any sentence that notifies its users that it uses cookies.

To increase the efficiency of the PDPC, your team is considering whether to adopt the abovementioned model to automate the analysis of data privacy policies. If this new method of analysis is adopted, the PDPC would rely entirely on the model's predictions to confirm whether app developers have breached the PDPA. The app developers would face a fine of up to $10,000 if they are found to have breached the PDPA.

How far would you, as a committee member of the PDPC:

50. Think that using the model is an effective method of identifying violations of the PDPA? *

Mark only one oval.

Not effective at all

1

2

3

4

5

Very effective

51. Think that using the model is a fair method of identifying violations of the PDPA? *

Mark only one oval.

Very unfair

1

2

3

4

5

Very fair

52. Think that using the model is a method that could be a risk to society? *

Mark only one oval.

Very risky

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Not risky

53. Trust the prediction made by the model? *

Mark only one oval.

Do not trust at all

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Trust very much

Context 3: Imagine that you are a user of an app. You read in a forum where other users allege that the app uses cookies. You decide to analyse the data privacy policy of the app using the abovementioned model and the model informs you that the data privacy policy does not contain any sentence that notifies its users that it uses cookies.

You are deciding whether to submit this prediction as the only supporting piece of evidence to the PDPC to claim that the app has used cookies without notifying you.

If the PDPC decides that the developer has indeed violated the PDPA, you could claim compensation from the app developer of up to $10,000.

How far would you, as a user of the app:

54. Think that using the model is an effective method of identifying violations of the PDPA? *

Mark only one oval.

Not effective at all

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Very effective

55. Think that using the model is a fair method of identifying violations of the PDPA? *

Mark only one oval.

Very unfair

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Very fair

56. Think that using the model is a method that could be a risk to society? *

Mark only one oval.

Very risky

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Not risky

57. Trust the prediction made by the model? *

Mark only one oval.

Do not trust at all

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Trust very much

B Summary of survey results

The `.csv` containing all the responses can be found here.

Survey: Using Explainable AI (XAI) Techniques on a Data Privacy dataset

31 responses

Publish analytics

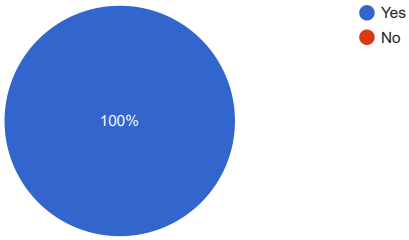
If you are a NUS / Yale-NUS student, are you 18 years old and above?

Copy

OR

If you are not a NUS student, are you above 21+ years old?

31 responses

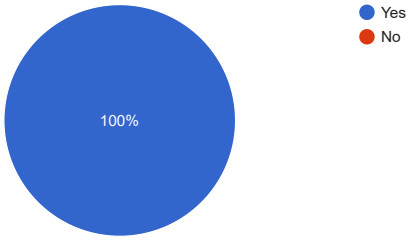


Participant Information Sheet

I have read about the purpose of this research study, agree to participate, and understand that I can withdraw at any time.

Copy

31 responses

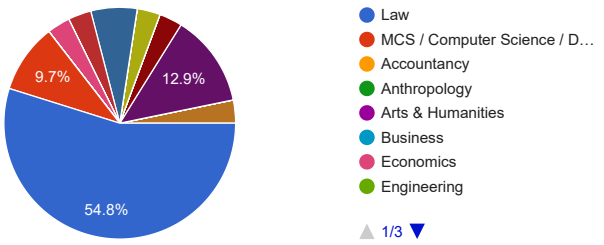


Part 1

What is your major / prospective major?

Copy

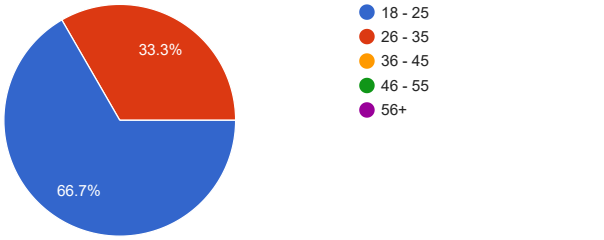
31 responses



What is your age?

Copy

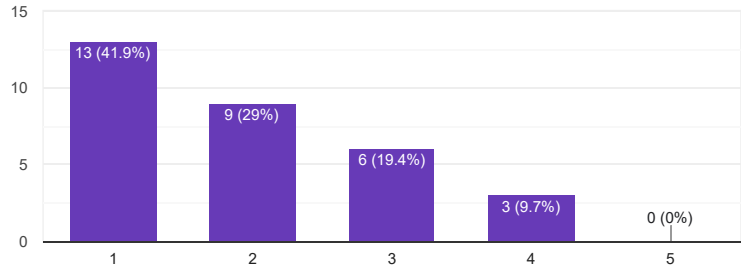
3 responses



Do you have any experience with AI / data science / programming?

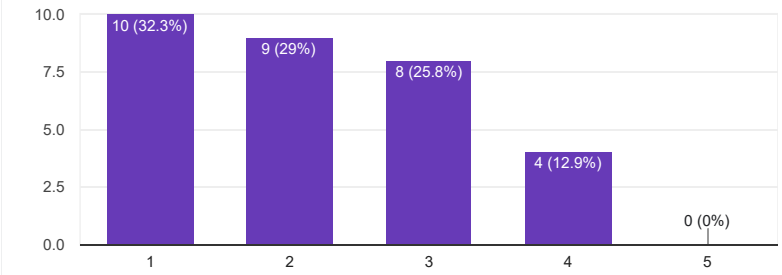
Copy

31 responses



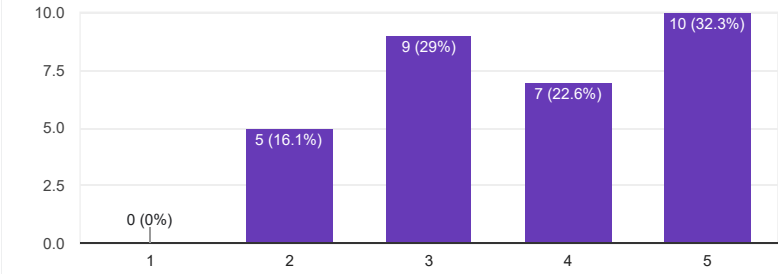
Do you have any experience with regarding data privacy / law? Copy

31 responses



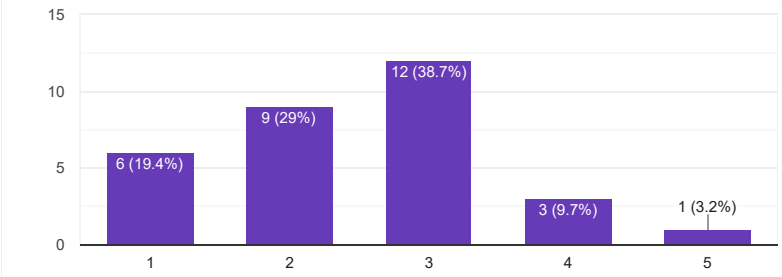
Are you concerned about your data privacy? Copy

31 responses



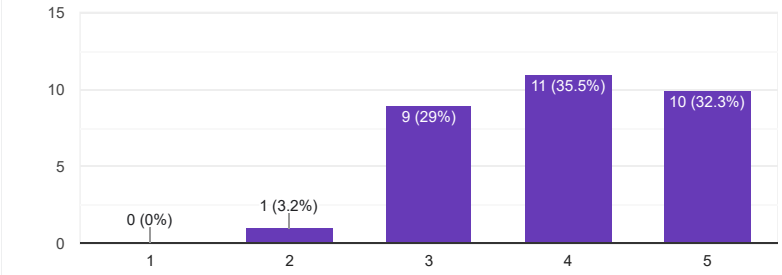
How would you rate your capability in protecting your online data? Copy

31 responses



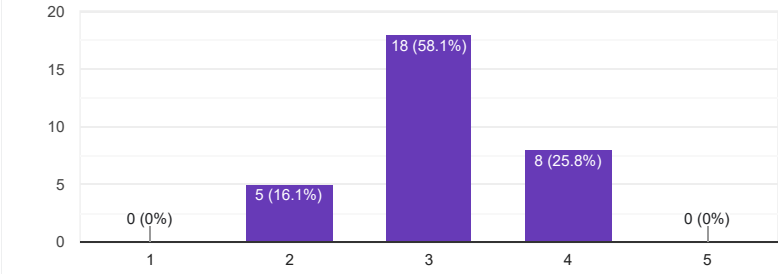
Do you think decisions that are made by AI can be useful to society? Copy

31 responses



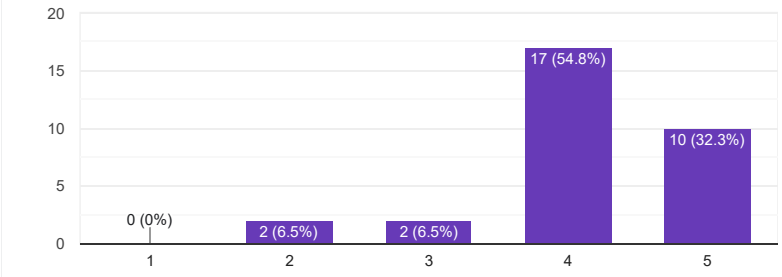
Do you think decisions made by AI are fair? Copy

31 responses



Do you think decisions made by AI can be a risk to society? Copy

31 responses

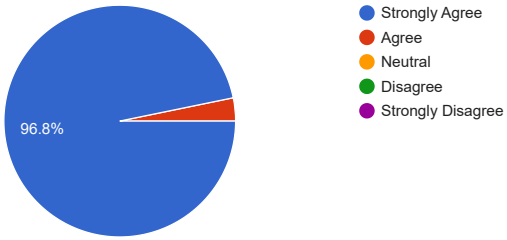


Part 2

Please select "strongly agree" to show that you are paying attention to this question.

Copy

31 responses



In this section, I will describe three different contexts with similar facts that relate to the use of the abovementioned model in analysing data privacy policies.

Each context corresponds with the perspective of an app developer, a member of the Personal Data Protection Commission (PDPC), and an user of the app.

I would then ask you questions to capture how your opinions on the use of AI in decision making would differ based on these three different perspectives.

Context 1: Imagine that you are an app developer. You are developing an app that uses cookies to track user activity online. To comply with the PDPA, you know that you need to include a sentence in your app's data privacy policy that notifies and asks for users' consent to use cookies.

Since you have no knowledge of the PDPA, you use the abovementioned model to analyse a pre-drafted data privacy policy that you found online. The model informs you that there is a sentence which states that cookies are being used.

You are deciding whether to rely entirely on the model's prediction, or pay costly legal fees to confirm with your friend who is a lawyer.

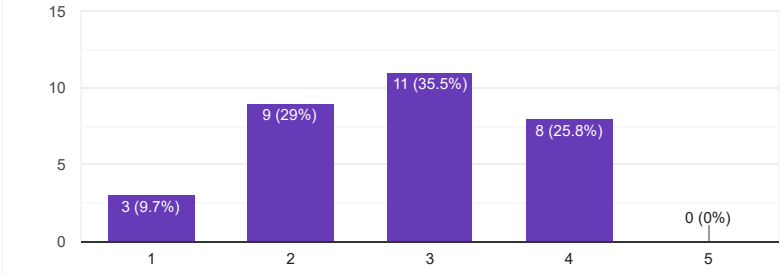
If the pre-drafted data privacy policy actually does not state that cookies are being used but your app uses cookies, you could face a fine of up to $10,000 in breach of the PDPA as you would have failed to notify your users.

How far do you, as the app developer:

Think that using the model is an effective method of identifying violations of the PDPA?

Copy

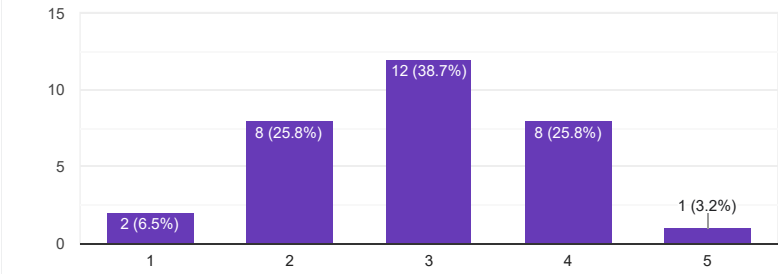
31 responses



Think that using the model is a fair method of identifying violations of the PDPA?

Copy

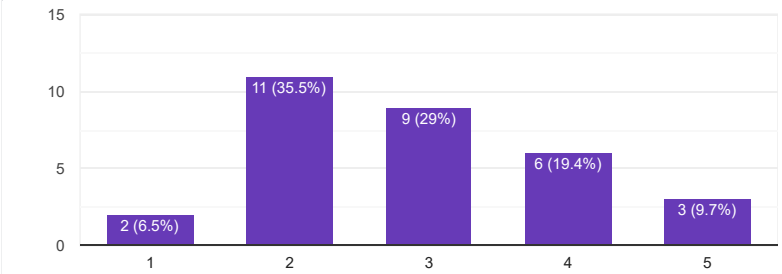
31 responses



Think that using the model is a method that could be a risk to society?

Copy

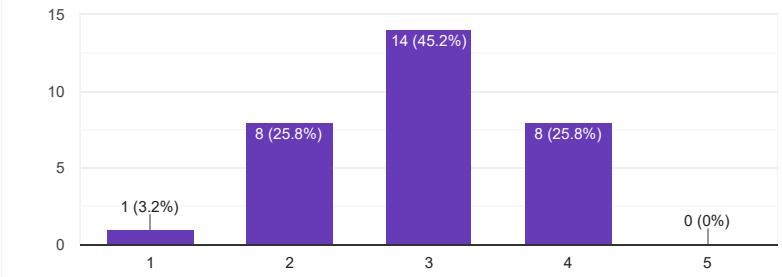
31 responses



Trust the prediction made by the model?

Copy

31 responses



Context 2: Imagine that you are a committee member part of the Personal Data Protection Commission (PDPC). A user of an app has informed you that an app is using cookies but has not notified its users.

Your team checks the code of the app and confirms that the app is indeed using cookies. Your team uses the abovementioned model and the model informs you that the data privacy policy does not contain any sentence that notifies its users that it uses cookies.

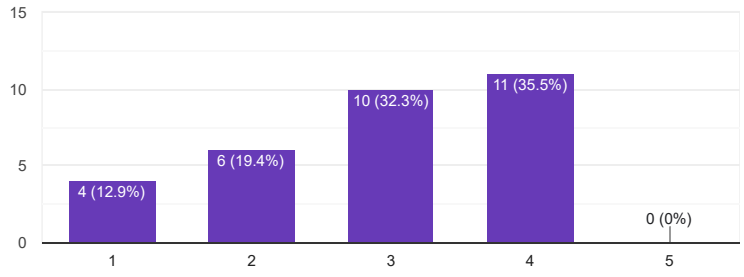
To increase the efficiency of the PDPC, your team is considering whether to adopt the abovementioned model to automate the analysis of data privacy policies. If this new method of analysis is adopted, the PDPC would rely entirely on the model's predictions to confirm whether app developers have breached the PDPA. The app developers would face a fine of up to $10,000 if they are found to have breached the PDPA.

How far would you, as a committee member of the PDPC:

Think that using the model is an effective method of identifying violations of the PDPA?

Copy

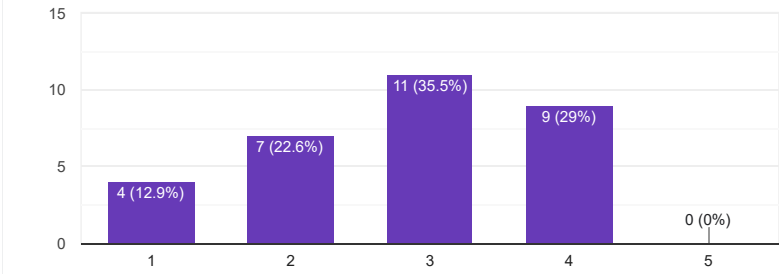
31 responses



Think that using the model is a fair method of identifying violations of the PDPA?

Copy

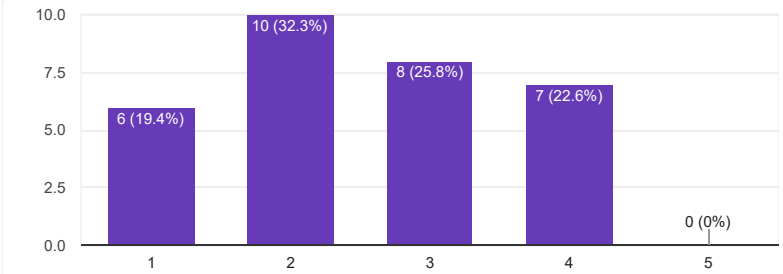
31 responses



Think that using the model is a method that could be a risk to society?

Copy

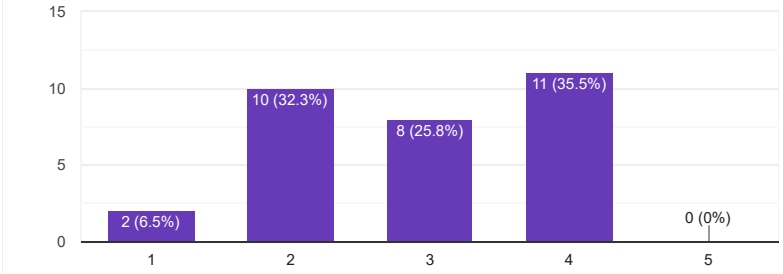
31 responses



Trust the prediction made by the model?

Copy

31 responses



Context 3: Imagine that you are a user of an app. You read in a forum where other users allege that the app uses cookies. You decide to analyse the data privacy policy of the app using the abovementioned model and the model informs you that the data privacy policy does not contain any sentence that notifies its users that it uses cookies.

You are deciding whether to submit this prediction as the only supporting piece of evidence to the PDPC to claim that the app has used cookies without notifying you.

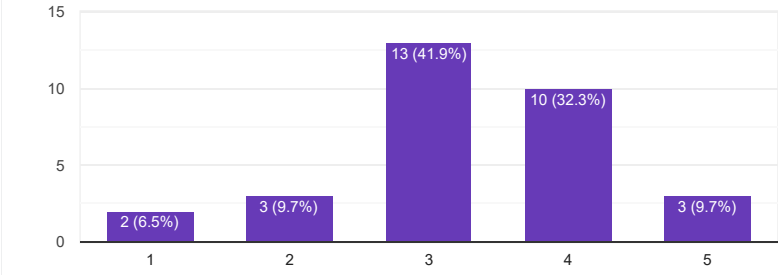
If the PDPC decides that the developer has indeed violated the PDPA, you could claim compensation from the app developer of up to $10,000.

How far would you, as a user of the app:

Think that using the model is an effective method of identifying violations of the PDPA?

Copy

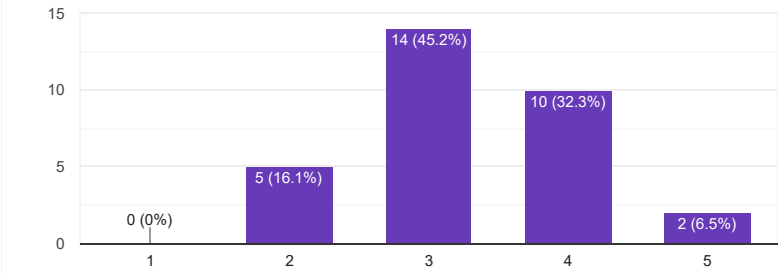
31 responses



Think that using the model is a fair method of identifying violations of the PDPA?

Copy

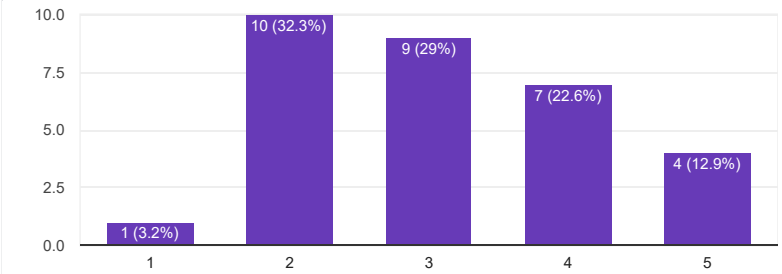
31 responses



Think that using the model is a method that could be a risk to society?

Copy

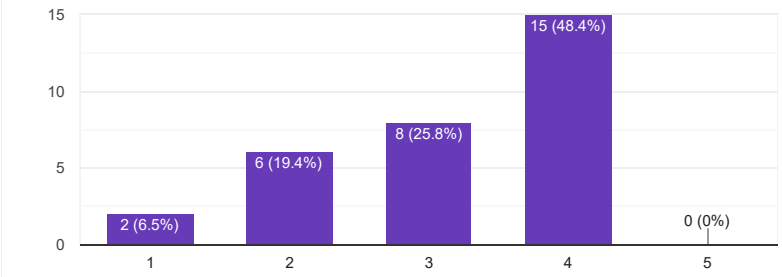
31 responses



Trust the prediction made by the model?

Copy

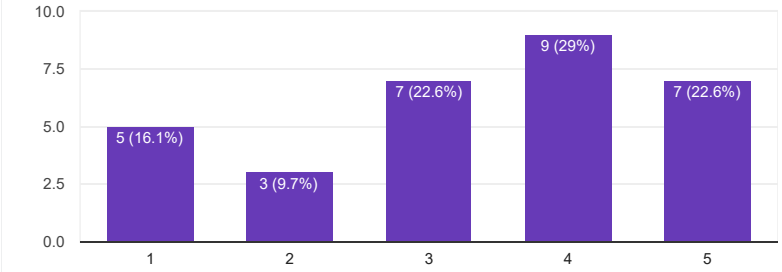
31 responses



Part 3

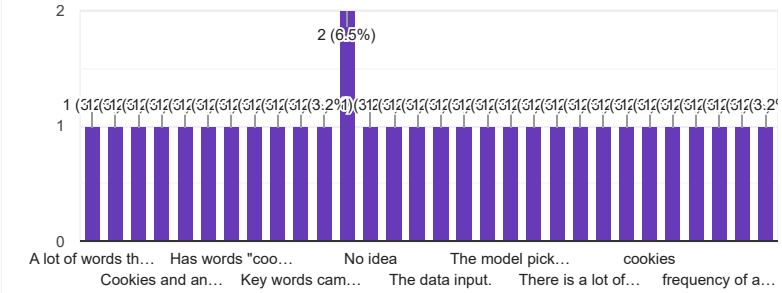
Do you understand why the model made the prediction? Copy

31 responses



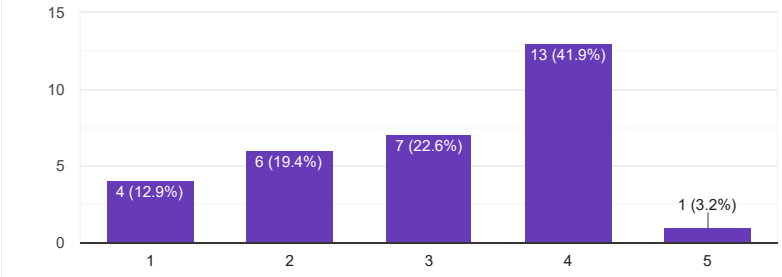
Why do you think the model made this prediction? Copy

31 responses



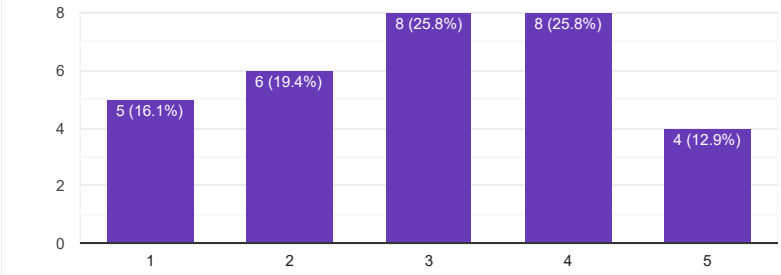
Did you find the visualisation easy to interpret? Copy

31 responses



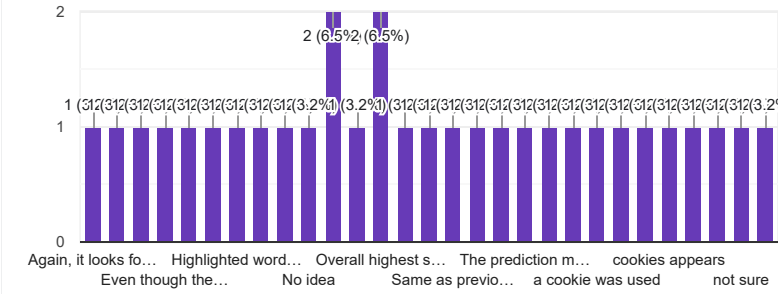
Do you understand why the model made the prediction? Copy

31 responses



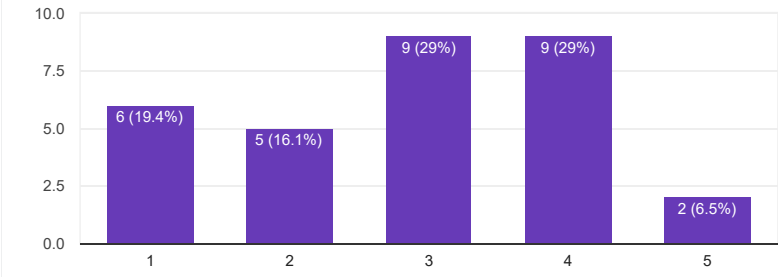
Why do you think the model made this prediction? Copy

31 responses



Did you find the visualisation easy to interpret? Copy

31 responses

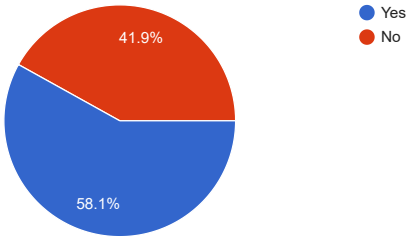


Based on your current understanding, do you think that the sentence below would be predicted to be "Identifier_Cookie_or_Similar_Tech_1stParty"?

"We also use **tracking technologies** to keep records, store your preferences, improve our advertising, and collect Non-Identifying Information, including Device Data and information about your interaction with the Site and our Business Partners' web sites."

Copy

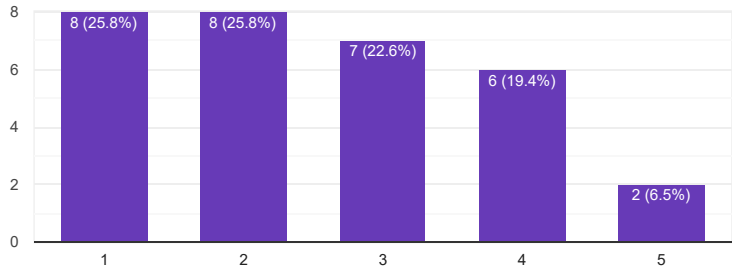
31 responses



Do you understand why the model made the prediction?

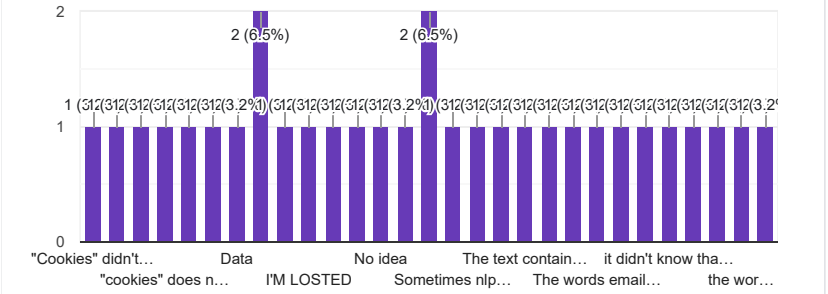
Copy

31 responses



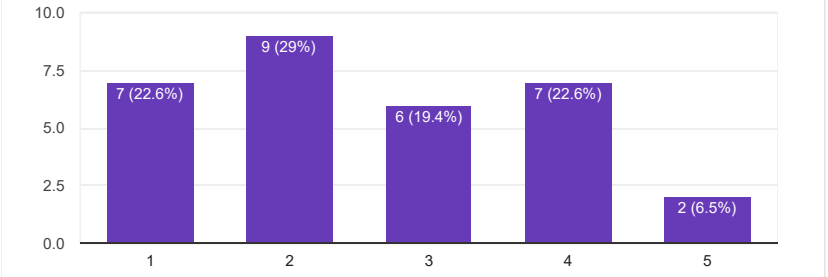
Why do you think the model made this prediction? Copy

31 responses



Did you find the visualisation easy to interpret? Copy

31 responses

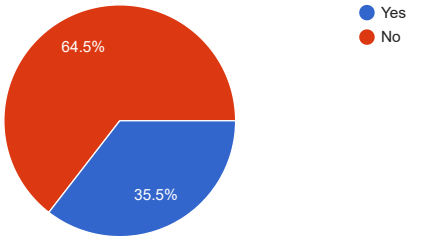


Based on your current understanding, do you think the sentence below would be predicted to be in "Identifier_Cookie_or_Similar_Tech_1stParty?"

"As explained above, you may either volunteer to us certain information (such as your **phone number**), or we may automatically collect certain information, such as through the use of your mobile device system's permissions, or through the use of cookies or similar tracking technologies."

Copy

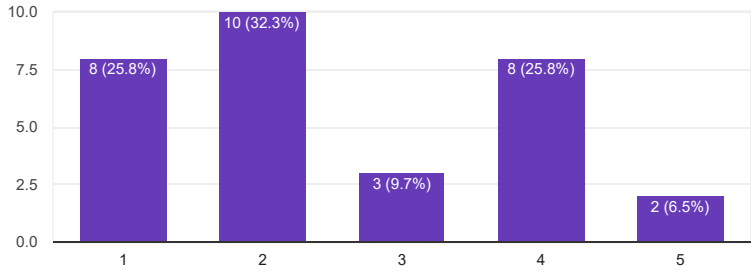
31 responses



Do you understand why the model made the prediction?

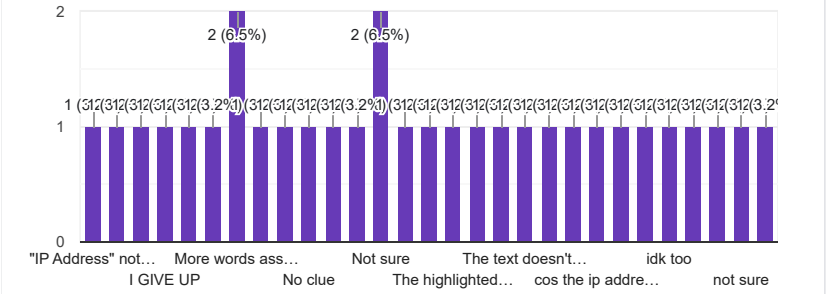
Copy

31 responses



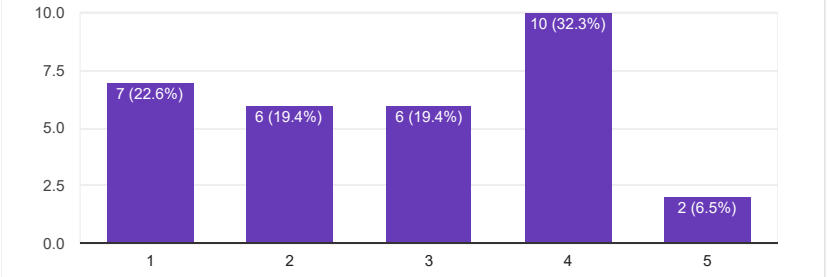
Why do you think the model made this prediction? Copy

31 responses



Did you find the visualisation easy to interpret? Copy

31 responses

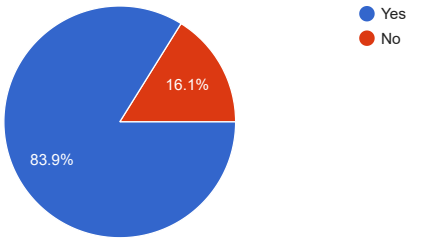


Based on your current understanding, do you think the sentence below would be predicted to be in "Identifier_Cookie_or_Similar_Tech_1stParty?"

"These **cookies and other such tracking technologies** allow the collection of data, such as your device's model, operating system and screen size, the other applications installed on your device, and information about how you use our services."

Copy

31 responses

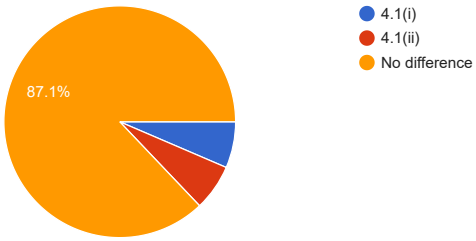


Part 4

Which explanation did you find easier to interpret?

Copy

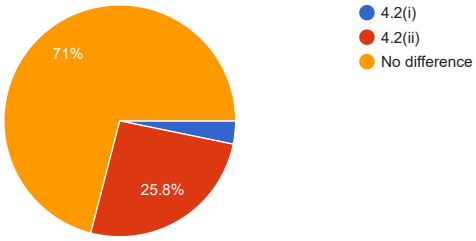
31 responses



Which explanation did you find easier to interpret?

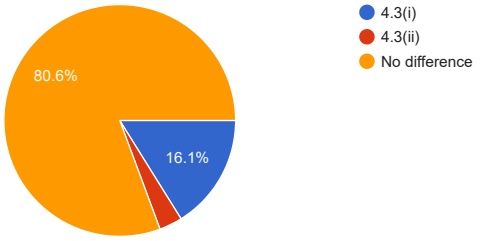
Copy

31 responses



Which explanation did you find easier to interpret? Copy

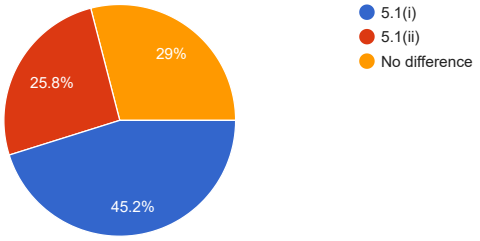
31 responses



Part 5

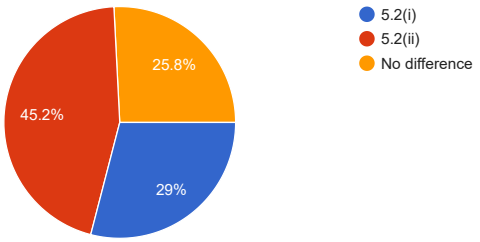
Which explanation did you find easier to interpret? Copy

31 responses



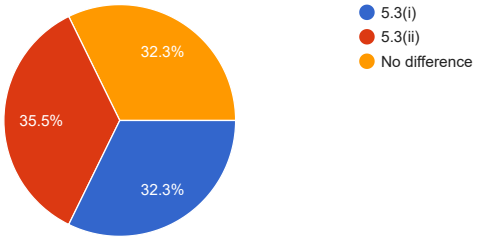
Which explanation did you find easier to interpret? Copy

31 responses



Which explanation did you find easier to interpret? Copy

31 responses



Part 6

Context 1: Imagine that you are an app developer. You are developing an app that uses cookies to track user activity online. To comply with the PDPA, you know that you need to include a sentence in your app's data privacy policy that notifies and asks for users' consent to use cookies.

Since you have no knowledge of the PDPA, you use the abovementioned model to analyse a pre-drafted data privacy policy that you found online. The model informs you that there is a sentence which states that cookies are being used.

You are deciding whether to rely entirely on the model's prediction, or pay costly legal fees to confirm with your friend who is a lawyer.

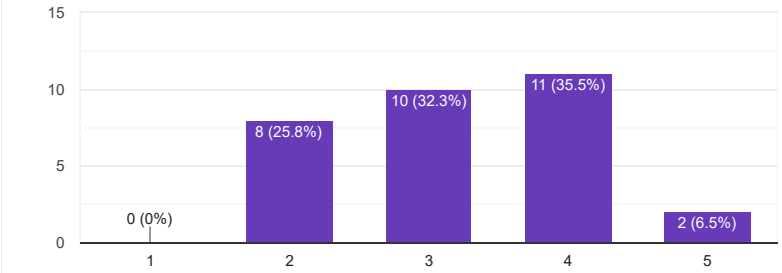
If the pre-drafted data privacy policy actually does not state that cookies are being used but your app uses cookies, you could face a fine of up to $10,000 in breach of the PDPA as you would have failed to notify your users.

How far do you, as the app developer:

Copy

Think that using the model is an effective method of identifying violations of the PDPA?

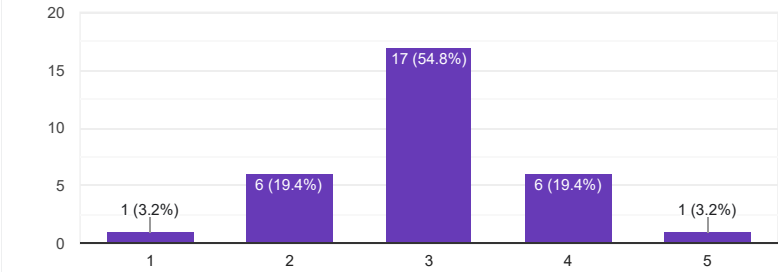
31 responses



Copy

Think that using the model is a fair method of identifying violations of the PDPA?

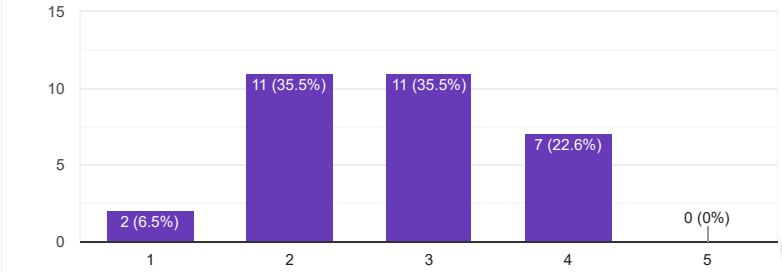
31 responses



Copy

Think that using the model is a method that could be a risk to society?

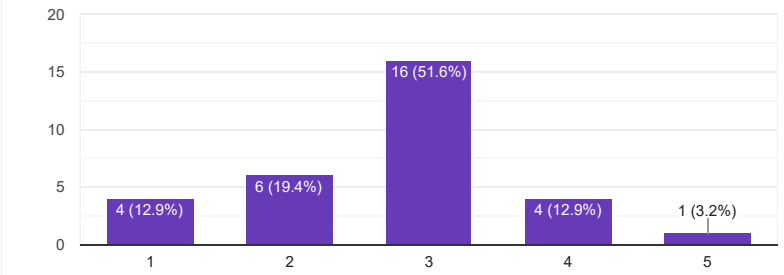
31 responses



Trust the prediction made by the model?

Copy

31 responses



Context 2: Imagine that you are a committee member part of the Personal Data Protection Commission (PDPC). A user of an app has informed you that an app is using cookies but has not notified its users.

Your team checks the code of the app and confirms that the app is indeed using cookies. Your team uses the abovementioned model and the model informs you that the data privacy policy does not contain any sentence that notifies its users that it uses cookies.

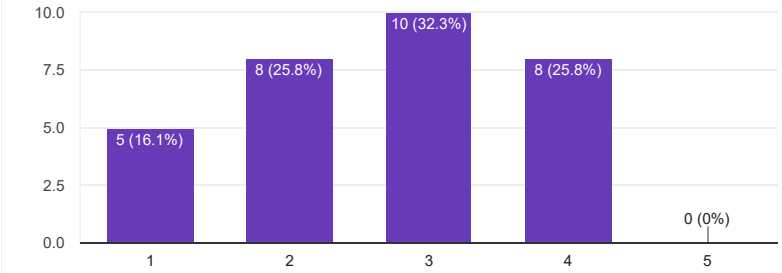
To increase the efficiency of the PDPC, your team is considering whether to adopt the abovementioned model to automate the analysis of data privacy policies. If this new method of analysis is adopted, the PDPC would rely entirely on the model's predictions to confirm whether app developers have breached the PDPA. The app developers would face a fine of up to $10,000 if they are found to have breached the PDPA.

How far would you, as a committee member of the PDPC:

Copy

Think that using the model is an effective method of identifying violations of the PDPA?

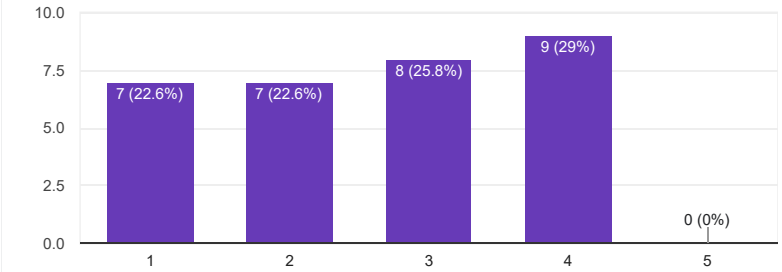
31 responses



Copy

Think that using the model is a fair method of identifying violations of the PDPA?

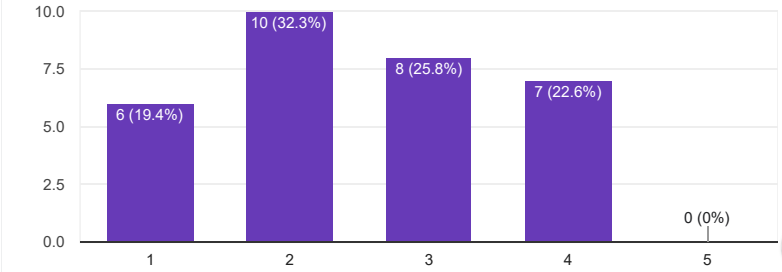
31 responses



Copy

Think that using the model is a method that could be a risk to society?

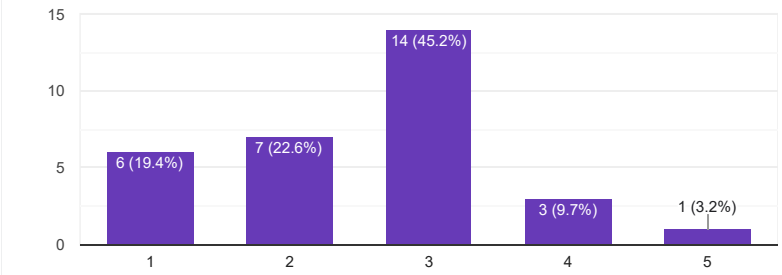
31 responses



Trust the prediction made by the model?

Copy

31 responses



Context 3: Imagine that you are a user of an app. You read in a forum where other users allege that the app uses cookies. You decide to analyse the data privacy policy of the app using the abovementioned model and the model informs you that the data privacy policy does not contain any sentence that notifies its users that it uses cookies.

You are deciding whether to submit this prediction as the only supporting piece of evidence to the PDPC to claim that the app has used cookies without notifying you.

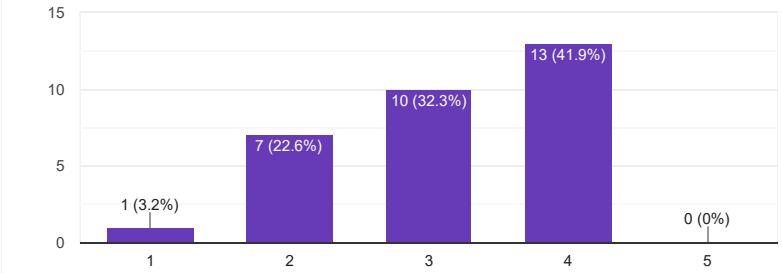
If the PDPC decides that the developer has indeed violated the PDPA, you could claim compensation from the app developer of up to $10,000.

How far would you, as a user of the app:

Copy

Think that using the model is an effective method of identifying violations of the PDPA?

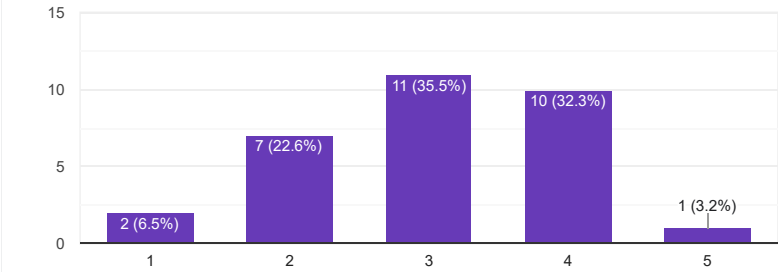
31 responses



Think that using the model is a fair method of identifying violations of the PDPA?

Copy

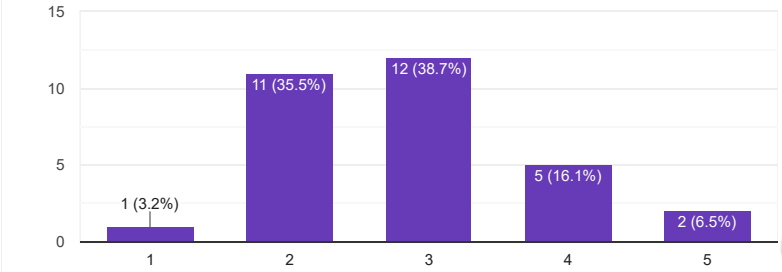
31 responses



Think that using the model is a method that could be a risk to society?

Copy

31 responses



Trust the prediction made by the model?

Copy

31 responses