



**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]

Unrestricted access

Make the Thesis immediately available for worldwide access.

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.

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.

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.

In recent years, the performance of machine learning models have increased but have come at the cost of decreased explainability. Machine learning has been gradually adopted in the legal industry to assist with low level legal analysis. This lack of explainability could be a significant hindrance towards greater adoption of artificial intelligence (AI) because the lawyer and law firm that use these models ultimately bear the responsibility of ensuring that their advice is legally accurate.

Separately, data privacy has been subjected to increased regulative oversight. With more ways to collect and use personal data, organisations are more likely to misuse such data. Therefore, users' awareness about how their data is collected and used should be heightened. Similarly, regulators and organisations have increased obligations to ensure that data is being collected and used responsibly.

This capstone uses this data privacy context to train reasonably performing models that are able to detect data practices within apps' data privacy practices. As the predictions of these models could be helpful to users, regulators and organisations, Explainable AI (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 current students and the general public about their self-reported levels of trust, fairness and perceptions of effectiveness of these visualisations.

“Any sufficiently advanced technology is indistinguishable from magic.”

Arthur C. Clarke

Contents

| | |
|---|-----|
| Acknowledgements | ii |
| Abstract | iii |
| 1 Introduction | 1 |
| 1.1 Motivation and significance | 1 |
| 1.1.1 The increasing opacity of NLP and the rise of XAI | 6 |
| 1.1.2 The importance of XAI in data privacy regulation | 9 |
| The consumer | 9 |
| The organisation | 11 |
| The state | 13 |
| 1.2 Problem statement | 16 |
| 1.3 Main findings and roadmap | 18 |
| 1.4 Font Formatting Commands | 18 |
| 1.4.1 Special characters | 18 |
| 2 Literature review and methodology | 20 |
| 2.1 Literature review | 20 |
| 2.1.1 Training models for NLP | 20 |
| 2.1.2 XAI methods for NLP | 20 |
| 2.1.3 Evaluating the effectiveness of XAI methods | 20 |
| User validation of explanations | 20 |

| | |
|--|----|
| 2.2 Methodology | 21 |
| 2.2.1 The APP-350 Corpus | 21 |
| 2.2.2 Data pre-processing | 23 |
| 2.2.3 Model training and XAI visualisation | 24 |
| 2.2.4 Proposed model training methodology | 25 |
| Text representation | 25 |
| Model choice | 26 |
| XAI method and package | 27 |
| 2.2.5 Tables | 28 |
| 2.2.6 Figures | 29 |
| 3 Exploratory Data Analysis (EDA) | 31 |
| 3.1 Sentence level | 31 |
| 3.2 Segment level | 32 |
| 4 Performance of models | 35 |
| 4.1 Models used and classification metrics | 35 |
| 4.1.1 Sentence level performance - TfIDF | 36 |
| 4.1.2 Sentence level performance - GloVe embeddings | 37 |
| 4.1.3 Segment level performance - TfIDF | 37 |
| 4.1.4 Segment level performance - GloVe embeddings | 37 |
| 4.2 Performance of individual classifiers for top 5 practices | 37 |
| 5 Discussion of survey results | 38 |
| 5.1 Summary of results | 38 |
| 5.1.1 Part 1: Beliefs relating to AI & data privacy | 39 |
| 5.1.2 Part 2 & Part 6: Comparison of self-reported scores of explainability across the three contexts | 39 |

| | | |
|-------|--|-----------|
| 5.1.3 | Part 3: Testing whether viewing more visualisations increase explainability | 44 |
| 5.1.4 | Part 4 & 5: Testing which model and word repre- sentation is more explainable | 44 |
| | Bibliography | 47 |
| | A Survey Questions | 52 |
| | B Summary of survey results | 91 |

List of Tables

| | | |
|-----|--|----|
| 2.1 | List of annotated data privacy practices and their descriptions. | 22 |
| 2.2 | The effects of treatments X and Y on the four groups studied. | 28 |
| 3.1 | Summary statistics for top 10 occurring practices at sentence level. | 31 |
| 3.2 | Summary statistics for bottom 10 frequently occurring practices. | 32 |
| 3.3 | Summary sentence statistics for top 10 frequently occurring practices. | 32 |
| 3.4 | Summary statistics of top 10 frequently occurring practices by segment. | 33 |
| 3.5 | Summary statistics for bottom 10 frequently occurring practices. | 33 |
| 3.6 | Summary segment statistics for top 10 frequently occurring practices. | 34 |
| 5.1 | Demographic breakdown of respondents according to academic discipline | 38 |
| 5.2 | p-values comparing whether there was a statistically significant increase / decrease in the explainability scores before and after viewing explanations. | 41 |

| | |
|---|----|
| 5.3 p-values comparing aggregated scores by context and metric | 41 |
| 5.4 p-values comparing aggregated scores across contexts and metrics | 41 |

List of Figures

| | |
|--|----|
| 2.1 An Electron | 30 |
| 5.1 Mean scores of self-reported beliefs of respondents regarding AI & data privacy. (1 = least agree, 5 = strongly agree. $n = 31$) | 40 |
| 5.2 Trend of the mean of self-reported understanding and interpretability after viewing each explanation | 43 |
| 5.3 Votes for predicting whether counterfactual would be classified correctly | 43 |
| 5.4 Testing for which model was more explainable: Respondents' votes to whether SVC + GloVe or Logistic regression + GloVe were more explainable | 45 |
| 5.5 Testing for which word representation was more explainable: Respondents' votes to whether SVC + TfIDF or SVC + GloVe were more explainable | 46 |

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 conversation with ChatGPT where I documented some of its legal writing abilities¹:

1. Explaining technical concepts.

Q: Explain what the eggshell skull rule means in negligence law.

A: The "eggshell skull" rule is a principle in tort law that states that a defendant is liable for all damages suffered by a plaintiff, even if the plaintiff's injuries are unusually severe or unexpected. The rule means that the defendant must take the plaintiff as they find them, meaning that they cannot argue that the plaintiff's injuries would not have been so severe if they were a "normal" person.

¹The question prompts and the names of the specific writing techniques were adapted from Bishop, 2023.

2. Adjusting for level and tone.

Q: Explain the eggshell skull rule like I'm five.

A: The "eggshell skull" rule means that if someone hurts another person, they are responsible for all the harm they cause, even if the person they hurt is more easily hurt than most people. It's like if you accidentally hit someone on the head and they have a very fragile skull, you would be responsible for all the damage, even if it's worse than if you hit someone with a normal skull. So you have to be extra careful not to hurt someone, even if they seem more delicate or breakable than most people.

3. Providing a legal opinion.

Q: I think that the eggshell skull rule is unfair. Can you make that sound academic?

A: Certainly. One could argue that the eggshell skull rule is inequitable, as it places an undue burden on defendants to anticipate and accommodate a plaintiff's pre-existing condition or vulnerability. By imposing liability on defendants for the full extent of the harm caused, regardless of the extent to which the harm would have been suffered by a normal person, the rule ignores the fact that the plaintiff's fragile condition is not the fault of the defendant. Additionally, the rule may incentivize plaintiffs to conceal pre-existing conditions, thereby undermining the efficiency of the tort system.

In 2017, some commentators opined that AI would be unlikely to take over higher-value roles of the lawyer such as advising clients, writing

legal briefs, negotiating and appearing in court (Lohr, 2017). However, as seen from the above, ChatGPT seems to have the ability to automate (or at least assist) these areas of legal practice, except perhaps negotiation and court appearances since ChatGPT does not have a physical presence. Indeed, most recently, ChatGPT has been said to have "the markings of an advocate" (Greene, 2022).

ChatGPT is the most recent AI model that uses advanced natural language processing (NLP). Natural language forms the bread and butter of the legal industry, as it is expressed in contracts, judgements and legislation. The legal industry has been adopting more machine learning tools to automate and assist low level legal analysis. Worldwide legal tech market revenues were at 27.6 billion USD and is projected to grow at a compound annual growth rate of 4% to 35.6 billion USD by 2027 (Statista, 2022). As early as 2018, LawGeex, a contract review startup, compared the performance of lawyers vs LawGeex's machine learning model in reviewing standard template Non-Disclosure Agreements (NDA). The model beat the humans both in terms of accuracy and time, with the model having a 94% accuracy rate and taking 26 seconds to complete the review. In comparison, the lawyers had an average accuracy of 85% and took 92 minutes to finish the task (LawGeex, 2018). More significantly, at the end of 2022, OpenAI's AI model GPT (Generative Pre-trained Transformer)² took the US bar examination and got 50% of the questions correct, and performed at a passing rate for both Evidence and Torts (Bommarito and Katz, 2022).

²The GPT is also used for ChatGPT.

Like the models that were trained by LawGeex and OpenAI, most legal tech tools that conduct low level legal analysis use natural language processing (NLP) techniques. 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:

Q: How did you think of that answer?

A: As an AI language model, I was programmed to provide responses based on a given prompt. In this case, I generated a response to your prompt based on my training data and my understanding of the English language.

ChatGPT does not seem to be able to explain its views like how a typical human would³. This lack of explainability could potentially be a significant hindrance towards NLP's further adoption within the legal industry because the lawyer and the law firm which use these models still ultimately bear the legal responsibility of ensuring that the analysis is legally sound. For example, r5(2)(c) of the Legal Profession (Professional Conduct) Rules 2015 states that a legal practitioner must act with reasonable diligence and competence in the provision of services to the client. A lawyer that relies on the analysis of legal tech tools and does

³Further prompting led ChatGPT to provide a list of academic papers that support its view.

not understand how the analysis was produced could be considered as 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. At the same time, explainable AI (XAI) techniques and research have been rising in popularity since 2020 (Linardatos, Papastefanopoulos, and Kotsiantis, 2020) but have not been specifically applied onto legal text. Therefore, this capstone aims to bridge the gap between the lawyer and the data scientist by using XAI techniques to explain the predictions of machine learning models.

Separately, the widespread collection and use of data by organisations in recent years has led to an increase of regulations governing data privacy. This "datafication" of society includes the "transformation of interactions into data that can be valued and used for predictive analysis". Governments have therefore stepped up their efforts to guarantee 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. A possible area of legal tech would be tools to aid in the compliance of these regulations. Therefore, I focus on NLP and XAI in the specific context of data privacy. This context provides a realistic evaluation of the interpretability of models that are trained on legal texts relating to data

privacy⁴.

1.1.1 The increasing opacity of NLP and the rise of XAI

In NLP, there are inherent difficulties in translating natural language to a mathematical representation that can be understood by the computer. Consider the sentence: "I am cow." Using a Bag-of-Words approach which represents words in terms of their frequency in the sentence, the word "cat" can be represented as:

$$\text{cow} = [0, 0, 1]$$

Therefore every word in this sentence can be encoded as:

$$\text{I} = [1, 0, 0]$$

$$\text{am} = [0, 1, 0]$$

$$\text{cow} = [0, 0, 1]$$

Each word can then be represented as a point with x, y, z coordinates in a three-dimensional space. The plane that passes through all three points would be a graphical representation of the sentence "I am cow". Mathematically, the equation $x + y + z - 1 = 0$ describes the sentence. Clearly there is a big difference in abstraction between the original sentence and the mathematical representation of the same sentence! Without knowing the process of how this equation was derived from the sentence, this equation could possibly refer to any other 3D object, such as the surface of a table.

Further, the Bag-of-Words approach only represents the sentence by the number of occurrences of each word in the sentence. The semantic

⁴All code and analysis used in this capstone can be found at <https://github.com/TristanKoh/capstone-repo/>.

difference between "I am cat" and "I am cow" (both are four legged animals that are mammals) vs "I am book" and "I am cow" (a cow is a living thing while a book is non-living) is not captured at all by the Bag-of-Words representation⁵. To the computer, "I am cat" and "I am book" are equally different from "I am cow" since both sentences respectively contain an instance of "cat" and "book" which "I am cow" does not contain. In fact, the sentence "I am cow cow" would make sense to the computer even though it makes no sense semantically.

Evidently, even the simplest word representation in NLP poses significant limitations in capturing how humans understand language. Word representations have since progressed tremendously in capturing more semantic meaning, as exemplified by GPT being able to pass certain subjects in the US 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).

Therefore, XAI for NLP has been designed to combat such issues. Explainability in this context is seen "from the perspective of an end user whose goal is to understand how a model arrives at its result" (Danilevsky et al., 2020). Danilvesky et al. categorises XAI for NLP according to two

⁵Furthermore, these two sentences cannot be represented graphically since it introduces a fourth dimension.

categories: explainability techniques and visualisation techniques. 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. In contrast, visualisation techniques are different ways to present these mathematical justifications to the end user. One example is a saliency heatmap, which highlights the combination of words / phrases that gave rise to the prediction at differing intensities⁶. (add figures here to show the differences) Overall, XAI aims to reduce the abstraction and opacity of NLP techniques so that end users can understand how the model arrived at a decision, which makes the "black-box" more of a "glass-box".

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". Interpretability is sometimes used to describe the model's internal logic and explainability as the ability of the user to understand that logic. In other cases, explainability is the attempt to explain the model's logic post-hoc without necessarily being representative of the model's true decision (Rosenfeld, 2021). Since arguing what should be the "correct" definition of explainability is out of the scope of this capstone, I use explainability and interpretability to refer to the general goal of XAI as stated above.

Another growing area of discourse is differentiating what explainability means to different users of the model who have different objectives

⁶More details about these XAI techniques would be covered later in Chapter 2.

for the explanations. For example, what could be explainable to data scientists may not be explainable to laypersons. While data scientists may find that providing more technical details about the model would make the model more explainable, laypersons might be more confused if too many details are provided. Further, if a person wants the model to explain why it recommended a certain product over another to purchase on Amazon, the type and level of detail of explanation would be very different compared to a situation where the model predicts that the accused has a higher risk of recidivism which therefore warrants a longer prison term (Rosenfeld, 2021). Therefore, assessing the effectiveness of XAI is highly dependent on the specific context and needs of the users. As will be explained in the following section, I focus specifically on what explainability means to three different stakeholders, the consumer, organisation and state for the purposes of analysing apps' data privacy policies.

1.1.2 The importance of XAI in data privacy regulation

As with any regulatory activity, there are three stakeholders: The consumer, the organisation and the state. The importance of XAI in data privacy can also be analysed from the perspectives of these three stakeholders.

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. Using NLP,

Wagner conducted a text analysis of privacy policies from 1996 - 2021, and made the following findings (Wagner, 2022):

1. The length of the average policy has doubled in the last 10 years, with the introduction of the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) leading to significant increases in length.
2. In 2021, the average Flesch Reading Ease of the policies is roughly the same as compared to academic articles such as the Harvard Law Review.
3. In 2021, 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 day⁷.
4. Policies have including more vague wording and invasive data privacy practices, such as implicit collection of location data, no specification as to whether data is anonymised, no meaningful choice given to consumers when there is a change of policy, and generic wording about security measures used to protect data.

The 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)⁸. Combined with the datafication of society, consumers have diminishing control over how their data is collected and used.

⁷ Assuming that a consumer visits 1462 unique websites each year.

⁸ Mantelero also discusses a different model of data regulation to combat the issues of the current informed consent paradigm, but that is out of the scope for this capstone.

In view of the unreadability of data privacy policies, legal tech that leverages on XAI can be useful to reduce this unreadability. Legal tech able to analyse and explain the implications of data privacy policies in simple terms to the consumer would greatly aid making these policies more accessible to the consumer, help in allowing them to make informed consent, and saves them time and money as compared to seeking legal advice.

The organisation

Though there are many data privacy regulations globally, I focus my analysis on the EU and Singapore since EU's GDPR has the widest scope of any data privacy regulation. I also chose Singapore as this author is based in Singapore. The GDPR is applicable to any personal data processing linked to the EU, either when the entity processing the personal data is established within the EU or when an entity outside the EU offers goods and services to people within the EU (Council of the European Union, 2016b). Therefore, the GDPR is not limited to EU's geographic boundaries and can potentially apply to any entity in any country. Organisations in the EU are subjected to a data subject's "right to explanation". Under Recital 71, data subjects have the right to obtain an explanation of a decision reached solely through automated processing (Council of the European Union, 2016c). This right could be said to be supported by Art. 22, where data subjects have the right not to be subject to a decision based solely on automated processing (Council of the European Union, 2016a). 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.

In Singapore, explainability of AI decision is found in the non-binding Model AI Governance Framework released by the Personal Data Protection Commission (PDPC) (Personal Data Protection Commission, 2020). While this framework is not specifically about data privacy, AI governance is closely related to data protection as the development of AI models is dependent on how personal data is governed. As part of guidance on the operations management of AI models, organisations are advised to provide explanations on how AI models are incorporated into the decision making process of the organisations so as to build understanding and trust with those stakeholders that use their products. Explainability also forms part of stakeholder communication. 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. In this regard, the Governance Framework also encompasses the GDPR's right to explanation.

Though there has been debate as to the scope of the GDPR's right to explanation, whether it is binding, and how it is balanced against other interests protected by the GDPR (Chesterman, 2021a), such legislation in addition to Singapore's AI Governance Framework that create guidance with regard to the explanability of automated decision making 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.

The state

At a more macro level, the state's concern when AI is used in legal decision making is maintaining the integrity of the judicial system and regulatory activities. Three challenges arise because of opacity when AI is used in legal decision making in a data privacy context (Chesterman, 2021b):

1. Inferior decision making as opacity reduces opportunities to identify wrongdoing or areas of reasoning which can be improved.

For example, TikTok's algorithm that recommends users' videos on their feed collects and aggregates user data to predict their interests. This algorithm is known to be much more aggressive compared to other social networks and has played a crucial role in its recent spike in popularity (Hern, 2022). It optimises for time spent viewing each video (including the time that the user returns to the same video) to recommend more videos that aim to keep the user scrolling on the app. While TikTok has disclosed the broad factors that the algorithm uses, the actual algorithm is much more nuanced as seen at how an internal company document describes recommending more "sad" content to increase watch time which could lead to videos involving self-harm (Smith, 2021). With opacity in the algorithm, users (and regulators) do not know how their personal data is being used to recommend them videos. This could

lead to adverse social consequences when the algorithm recommends controversial content such as religiously or politically partisan posts. Opacity also prevents regulators from pre-empting or intervening in such undesirable social consequences such as mandating that such recommendation algorithms must have limits to the type of content that can be recommended.

2. Impermissible decisions such as reifying discrimination as opacity masks such decisions with the "objective" nature of AI predictions.

A common example of this is using AI to predict which bank customers would default on their loans. AI models are trained to predict the risk of default based on demographic metrics, such as employment status and income level. If the model assesses the customer is likely to default, the bank would probably not allow the customer to take out a loan. In some jurisdictions, anti-discrimination legislation prevents banks from making such assessments with regard to race, gender, colour, religion etc. which are considered personal data under data privacy regulation⁹. If opaque AI models are used, it would not be possible to ascertain whether these protected personal data are actually used for the prediction. It could be possible to justify a decision not to grant a loan just because the model made a prediction that the customer is likely to default. This prediction could be seen as "objective" as it was made by an algorithm rather than a human even though it actually perpetuates racism because it used race as a predictor (for example). Hence, opacity in AI

⁹Such as the Fair Housing Act in the US that makes it unlawful for any lender to discriminate against any person because of race, colour, religion, national origin, sex, handicap or familial status (Board of Governors of the Federal Reserve System, 2022).

models potentially pose societal and regulatory issues.

3. Legitimacy of decisions since the legal system depends as much on the justification of decisions as it does on the decisions themselves.

In the context of data privacy regulation in Singapore, Parliament enacts the PDPA while the PDPC is empowered to create regulations and enforce the PDPA¹⁰. There is an appeal process for cases appearing before the PDPC and the higher courts can overturn the PDPC's decisions¹¹. 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 court judgements, if there was no explanation of the model's judgement, there would be no basis for the higher courts to overturn judgements made by lower courts.

Further, in common law systems, judges' decisions become part of case law and past decisions become binding on new cases with the same facts (i.e. the doctrine of *stare decisis*). One of the benefits of case law is that it can be developed incrementally and adapted to suit real facts unlike legislation, so that the law actually meets the needs of the parties. While *stare decisis* does not apply to PDPC cases at first instance, the same principle of incrementally developing and applying the law still applies. For example, the quantum of fine imposed on organisations in breach of the PDPA is determined according to how severe the breach is in comparison with similar

¹⁰Per s65(1) of the PDPA.

¹¹Per s48Q, s48R and s54 of the PDPA.

past cases¹². If opaque AI is used to write or aid in the writing of these decisions, there would be fewer opportunities to improve the law since the reasoning would likely to be insufficiently explicit. Hence the development of the law would also be impeded.

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. XAI also provides the state with more transparency in regulating organisations' use of AI and aiding in judicial decision making.

1.2 Problem statement

The foregoing analysis of the issues in data privacy and the opacity of NLP models can be summarised as follows: 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. More research has been put into XAI in an effort to combat this opacity, but it is an open question as to whether the explanations are effective to the different users and use cases of the models. Data privacy regulation can benefit from the use of more XAI in automated decision making. Thus, there is a gap in research where XAI is assessed and applied in data privacy.

Hence, I choose to train AI models that are able to identify data privacy practices in apps' data privacy policies. A privacy practice describes

¹²An example would be *RedMart Limited* [2022] SGPDPC 8 at [33].

a certain behaviour of an app that can have privacy implications¹³. In this regard, these models function similarly to that of LawGeex's contract review models which are able to identify the type of clause, such as which party owns the intellectual property arising from the performance of the contract¹⁴. XAI methods are then used to visualise why the models made such predictions. Such a ;legal tech tool saves consumers time by highlighting specific data privacy practices upfront so that they can make informed consent.

I focus on these specific problems for the rest of the capstone:

1. Which AI models perform the best on these privacy policy texts and why?
2. Which models are the most explainable to an audience that includes laypersons and people with domain knowledge in data science and the law, and why?
3. Are there differences in how people perceive explainability if they were asked to consider the predictions of these models from the perspective of a consumer, an organisation or the PDPC (as the regulatory authority for data privacy), and why?

¹³This will be further elaborated on when **the dataset** is introduced.

¹⁴Automated contract review models are also able to analyse the level of legal risk associated with any amendments made to the clause, such as if the counterparty amended the wording of the clause to state that only the counterparty owns the intellectual property instead of both contracting parties.

1.3 Main findings and roadmap

Chapter 2 is a literature review of XAI methods of NLP and methods of assessing these XAI methods which forms the justification of the methodology and dataset used. Chapter 3 presents an exploratory data analysis of the dataset. Chapter 4 reports the performance of the models that were trained on the dataset.

1.4 Font Formatting Commands

Similarly to Word, LaTeX provides simple formatting, including **bold**, *italic*, underlined and ugly stuff. However, no underline or strikethrough by default. You can also change the size of the text, using , , , **huge**.

1.4.1 Special characters

In that case, simply use \$ (by the way, note that using the dollar sign in your text switches to mathematical notation. To actually print a dollar sign use the \textdollar command). The equation above has a label, meaning you can refer to it. The numbering system uses the chapter number (in this case 1), then the equation position within the chapter (1 again). Example: Equation ?? is an example of an equation in LaTeX. In case you would like to have an equation without numbering it? Easy!

$$t = a \times \log_2\left(\frac{D}{W} + 1\right) + b$$

The only difference? The * symbol in the \begin{equation*}. This also works with Figures and Tables.

2 Literature review and methodology

We can reference other chapters, for example, here we refer to Chapter 1.

2.1 Literature review

2.1.1 Training models for NLP

(Katz et al., 2023)

2.1.2 XAI methods for NLP

Welcome to this L^AT_EX Thesis Template, a beautiful and easy to use template for writing a thesis using the L^AT_EX typesetting system.

2.1.3 Evaluating the effectiveness of XAI methods

(Doshi-Velez and Kim, 2017)

User validation of explanations

(Górski and Ramakrishna, 2021) (Rosenfeld, 2021)

2.2 Methodology

There are three major steps to the capstone: First is data pre-processing, second is model training and applying XAI techniques to visualise their predictions ("Model Training and XAI visualisation"), and the last is to survey law and non-law students about whether they find these explanations interpretable ("Survey to assess interpretability"). I describe the specific methodology of these parts below.

2.2.1 The APP-350 Corpus

The APP-350 Corpus consists of 350 annotated Android app privacy policies. The corpus has been used by a previous paper to train models to detect data privacy practices (Zimmeck et al., 2019). Each annotation consists of a practice and a modality. A "privacy practice" (or "practice") 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).

As not all practices had modalities, altogether, 57 different categories were annotated. The following is a table of the practices and their descriptions.

The APP-350 Corpus was used in a broader project to train machine learning models to conduct a privacy census of 1,035,853 Android apps. In that project, the researchers downloaded the data privacy practices of all apps from the Play Store with more than 350 million installs (which

| Data Type | Description |
|-----------------------------------|--|
| Contact | The policy describes collection of unspecified contact data. |
| Contact_Address_Book | The policy describes collection of contact data from a user's address book on the phone. |
| Contact_City | The policy describes collection of the user's city. |
| Contact_E_Mail_Address | The policy describes collection of the user's e-mail. |
| Contact_Password | The policy describes collection of the user's password. |
| Contact_Phone_Number | The policy describes collection of the user's phone number. |
| Contact_Postal_Address | The policy describes collection of the user's postal address. |
| Contact_ZIP | The policy describes collection of the user's ZIP code. |
| Demographic | The policy describes collection of the user's unspecified demographic data. |
| Demographic_Age | The policy describes collection of the user's age (including birth date and age range). |
| Demographic_Gender | The policy describes collection of the user's gender. |
| Identifier | The policy describes collection of the user's unspecified identifiers. |
| Identifier_Ad_ID | The policy describes collection of the user's ad ID (such as the Google Ad ID). |
| Identifier_Cookie_or_similar_Tech | The policy describes collection of the user's HTTP cookies, flash cookies, pixel tags, or similar identifiers. |
| Identifier_Device_ID | The policy describes collection of the user's device ID (such as the Android ID). |
| Identifier_IMEI | The policy describes collection of the user's IMEI (International Mobile Equipment Identity). |
| Identifier_IMSI | The policy describes collection of the user's IMSI (International Mobile Subscriber Identity). |
| Identifier_IP_Address | The policy describes collection of the user's IP address. |
| Identifier_MAC | The policy describes collection of the user's MAC address. |
| Identifier_Mobile_Carrier | The policy describes collection of the user's mobile carrier name or other mobile carrier identifier. |
| Identifier_SIM_Serial | The policy describes collection of the user's SIM serial number. |
| Identifier_SSID_BSSID | The policy describes collection of the user's SSID or BSSID. |
| Location | The policy describes collection of the user's unspecified location data. |
| Location_Bluetooth | The policy describes collection of the user's Bluetooth location data. |
| Location_Cell_Tower | The policy describes collection of the user's cell tower location data. |
| Location_GPS | The policy describes collection of the user's GPS location data. |
| Location_IP_Address | The policy describes collection of the user's IP location data. |
| Location_WiFi | The policy describes collection of the user's WiFi location data. |
| SSO | The policy describes receiving data from an unspecified single sign on service. |
| Facebook_SSO | The policy describes receiving data from the Facebook single sign on service. |

TABLE 2.1: List of annotated data privacy practices and their descriptions.

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 $\alpha = 0.78^1$, the agreement between the annotations exceeded previous similar research.

Since the focus of this capstone is to assess the interpretability of XAI models specifically within a legal context, this dataset was chosen for the following reasons:

1. APP-350 contains real-world data privacy practices as they were scraped from Google PlayStore apps. Thus training XAI models on

¹Krippendorff's α is a measure of agreement, with $\alpha > 0.8$ indicating good agreement, $0.67 \leq \alpha \leq 0.8$ indicating fair agreement, and $\alpha < 0.67$ indicating doubtful agreement.

such a dataset would provide a realistic insight into the extent of which AI models are explainable in the legal context.

2. Legal tech companies are also using such datasets to train models as part of their contract / document review products. By using APP-350 to train XAI models, the results can be used as a (simple)² proxy for the explainability of models that are currently used in the industry.
3. APP-350 is a labelled dataset, allowing easy validation of results. If an unlabelled dataset was used, unsupervised training would have to be conducted. The performance of the 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 law specific labelled datasets to begin with.
4. APP-350 is labelled on both the sentence and segment (i.e. paragraph) level. This provides more granular data for training the AI models.

2.2.2 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

²The datasets used in industry are usually much larger and the models used are more complicated. However, APP-350 would be sufficiently complicated to serve as a toy example at an undergraduate level.

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.

2.2.3 Model training and XAI visualisation

As the original researchers used the same dataset to train classifiers to predict on unseen data privacy policies, I adopt their training methodology and model choice as a guide for this capstone.

The original researchers feature engineered the data as follows (Page 71 to 72):

1. Tokenisation: Lowercase all characters, remove non-ASCII characters, no stemming, normalisation of whitespace and punctuation, unigrams and bigrams.
2. Word representation: Union of TF-IDF and manually crafted features. The manually crafted features consist of Boolean values indicating the presence or absence of indicative strings the researchers observed in the data.

Individual classifiers were then trained for every policy classification. For all the classifications (except for four categories), they trained a model using the scikit-learn SVC 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.

(Add performance of researchers' models here.)

2.2.4 Proposed model training methodology

There are three main components to the model training methodology: Text representation, ML model, and XAI package used to explain the trained model.

Text representation

Computers cannot understand text directly and have to be converted into some kind of quantitative data. Therefore in NLP, text representations are methods to represent text as numeric or continuous vectors. This step is done before the model is trained. I use Tf-IDF (term frequency - inverse document frequency) and GloVe word embeddings as the text representations.

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.
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 would not be able to capture the semantic relationship between words. Word embeddings try to overcome this issue with count-based metrics like Tf-IDF. Word embeddings are vector representation of words, such that the vectors closer to each other in a vector space are similar in their semantic meaning.³

GloVe is one type of word embeddings (more information to be added)

Model choice

As the researchers found that the SVC classifier produces the best performance, I use SVC as well. In all, I use the following models:

1. Logistic regression
2. SVC
3. Ensemble classifiers (AdaBoost, GradientBoost, Random Forest)

³<https://becominghuman.ai/mathematical-introduction-to-glove-word-embedding-60f24154e54c>

Logistic regression functions as a baseline classifier for its simplicity. SGDClassifier functions a possible alternative to SVC since the SGDClassifier can use a linear SVM loss function. Ensemble classifiers are included to provide a wider range of models to compare performance.

The two broad types of AI models that are usually used are classical machine learning models (such as logistic regression and tree-based classifiers) and neural networks. Though recent advances in NLP are in the field of neural networks ⁴, I chose to focus only on classical ML models to reduce the possible complexity of the capstone, as the field of NLP by itself produces models that are usually more complex than models trained on quantitative variables. Explaining neural networks used for NLP would be a much more complex task compared to explaining classical ML models.

Further, as the APP-350 corpus only contains (insert number here) datapoints, there is insufficient data to train neural networks. Generally, neural networks require (add number here) datapoints to perform well.

XAI method and package

Local Interpretable Model-agnostic Explanations (LIME) is used in this capstone because it can be fitted to any model and is one of the few XAI packages that is easily implementable.

(some description of how LIME works here)

⁴State of the art NLP models include BERT and ELMo.

TABLE 2.2: The effects of treatments X and Y on the four groups studied.

| Groups | Treatment X | Treatment Y |
|---------------|--------------------|--------------------|
| 1 | 0.2 | 0.8 |
| 2 | 0.17 | 0.7 |
| 3 | 0.24 | 0.75 |
| 4 | 0.68 | 0.3 |

2.2.5 Tables

Tables are an important way of displaying your results, below is an example table which was generated with this code:

```
\begin{table}
  \caption{The effects of treatments X and Y on the four groups studied.}
  \label{tab:treatments}
  \centering
  \begin{tabular}{l l l}
    \toprule
    \tabhead{Groups} & \tabhead{Treatment X} & \tabhead{Treatment Y} \\
    \midrule
    1 & 0.2 & 0.8 \\
    2 & 0.17 & 0.7 \\
    3 & 0.24 & 0.75 \\
    4 & 0.68 & 0.3 \\
    \bottomrule
  \end{tabular}
\end{table}
```

You can reference tables with `\ref{<label>}` where the label is defined within the table environment. See `Chapter1.tex` for an example of

the label and citation (e.g. Table 2.2).

2.2.6 Figures

There will hopefully be many figures in your thesis (that should be placed in the *Figures* folder). The way to insert figures into your thesis is to use a code template like this:

```
\begin{figure}
\centering
\includegraphics{Figures/Electron}
\decoRule
\caption[An Electron]{An electron (artist's impression).}
\label{fig:Electron}
\end{figure}
```

Also look in the source file. Putting this code into the source file produces the picture of the electron that you can see in the figure below.

Sometimes figures don't always appear where you write them in the source. The placement depends on how much space there is on the page for the figure. Sometimes there is not enough room to fit a figure directly where it should go (in relation to the text) and so L^AT_EX puts it at the top of the next page. Positioning figures is the job of L^AT_EX and so you should only worry about making them look good!

Figures usually should have captions just in case you need to refer to them (such as in Figure 2.1). The \caption command contains two parts, the first part, inside the square brackets is the title that will appear in the *List of Figures*, and so should be short. The second part in the curly brackets should contain the longer and more descriptive caption text.



FIGURE 2.1: An electron (artist's impression).

The \decoRule command is optional and simply puts an aesthetic horizontal line below the image. If you do this for one image, do it for all of them.

L^AT_EX is capable of using images in pdf, jpg and png format.

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. The table below shows some summary statistics of the top 10 and bottom 10 frequently occurring practices at the 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_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.

| 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.

| | 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.

According to Table 4 below, 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.

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. The table below 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 7 below, 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

As seen in Table 2 and Table 5 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 models for the top N (where $3 \leq N \leq 10$) frequently occurring practices at both the sentence and segment level.

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.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Generally, precision is the preferred metric when the cost of false positives are high, such as detecting spam. If a classifier classifies a non-spam email as spam, the user would lose important information. Whereas recall is preferred when the costs of false negatives are high. For example, if a classifier classifies someone as not having cancer when they actually have cancer, the patient would lose the opportunity for early intervention. F1 is a harmonic mean of precision and recall, and is usually used to find a balance between precision and recall. Since there is neither a high cost for false positives or false negatives, F1 score is primarily used to assess the performance across the top N practices. As the distributions of records across the practices are uneven, I focus on the weighted average of F1 scores.

I also assess the top N performance for both the Tf-IDF and GLoVe word embeddings.

4.1.1 Sentence level performance - TfIDF

Generally we see that the SVC classifier performance the best across the metrics and across the top N frequently occurring practices. This corresponds with the findings by the researchers as they also found that the SVC classifier produced the best performance.

(add figures here)

4.1.2 Sentence level performance - GloVe embeddings

4.1.3 Segment level performance - TfIDF

4.1.4 Segment level performance - GloVe embeddings

4.2 Performance of individual classifiers for top 5 practices

5 Discussion of survey results

5.1 Summary of results

In total, 31 responses were collected. The table 5.1 below show the demographic data and aggregated responses relating to respondents' beliefs regarding AI and data privacy. While I originally intended to conduct analysis of the results across different demographics, there were limited non-law respondents. Therefore, I segmented the analysis that follows according to law (58%) vs non-law (42%) respondents.

| 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

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 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. Such a view seems optimistic, but also realistic.

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 $\alpha = 0.1$:

H_0 : There is no increase / decrease in scores after viewing the explanations.

H_1 : There is an increase / decrease in scores after viewing the explanations.

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 it 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 $d < 0$, then $i < f$ and

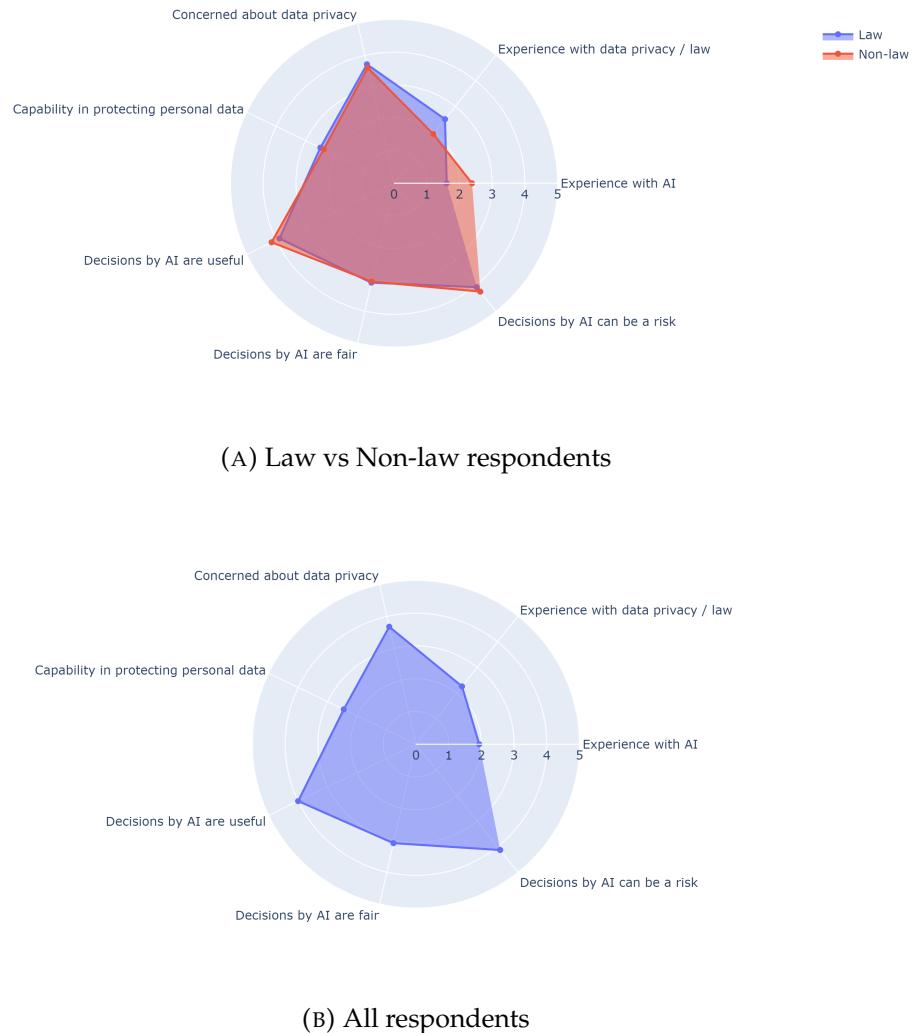


FIGURE 5.1: Mean scores of self-reported beliefs of respondents regarding AI & data privacy. (1 = least agree, 5 = strongly agree. $n = 31$)

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

The p-values are reported in Table 5.2 and ?? . "Increase" and "decrease" refer to the p-values of the test to check whether the final scores increased or decreased. Here are some observations:

| 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 pre- diction 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 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 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

1. There is a statistically significant decrease generally when comparing the averaged scores across all metrics and contexts. H_0 can be rejected in favour of H_1 . Thus self-reported overall explanability decreased after viewing the explanations.
2. Explanability for context 2 and 3 had a statistically significant decrease, when comparing the averaged metrics.
3. Trust was the only metric that statistically significantly decreased when comparing the metrics averaged by context.
4. There is a statistically significant decrease for effectiveness and trust in context 2, and fairness in context 3. The only statistically significant increase was for effectiveness in context 1, as well as fairness in context 3. For these tests, H_0 is rejected in favour of H_1 . Thus it can be stated that for these metrics, there was an increase / decrease after viewing the explanations.
5. While there is a statistically significant increase in effectiveness, trust decreased in context 1. In comparison, effectiveness and trust both significantly decreased in context 2. Hence there seems to be an inverse correlation between effectiveness and trust in context 1, and positive correlation in context 2.

Here are some inferences that can be drawn from these observations:

- 1.

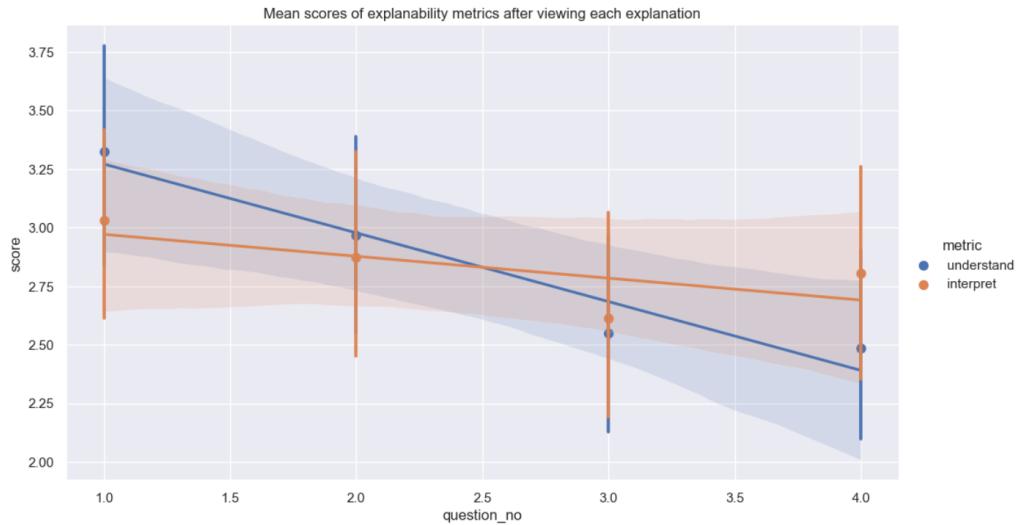


FIGURE 5.2: Trend of the mean of self-reported understanding and interpretability after viewing each explanation

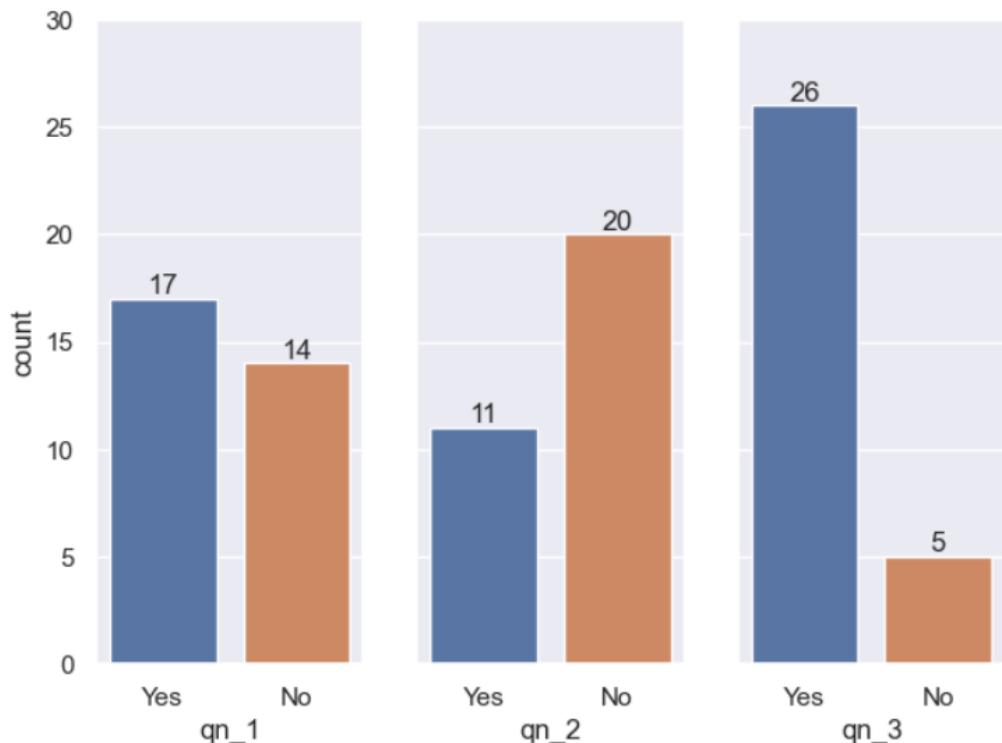


FIGURE 5.3: Votes for predicting whether counterfactual would be classified correctly

5.1.3 Part 3: Testing whether viewing more visualisations increase explainability

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

Overall, respondents found no difference between logistic regression and SVC (Figure 5.4), while it was more contentious for the word representation, with a third split across the three categories (Figure 5.5). For logistic regression vs SVC, the results coincided with the performance of the classifiers. However for word representations , I found GloVe more explainable than Tf-IDF. While there is no majority consensus in favour of GloVe, 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. This means that while respondents might not be able to accurately discern which was the more explainable word representation as compared to an expert in the field, they still have some level of intuition when there are differences in the components of the models used. This suggests that LIME used in this context has a decent level of explainability, if not the distribution of votes for Part 4 & Part 5 would be roughly equivalent.

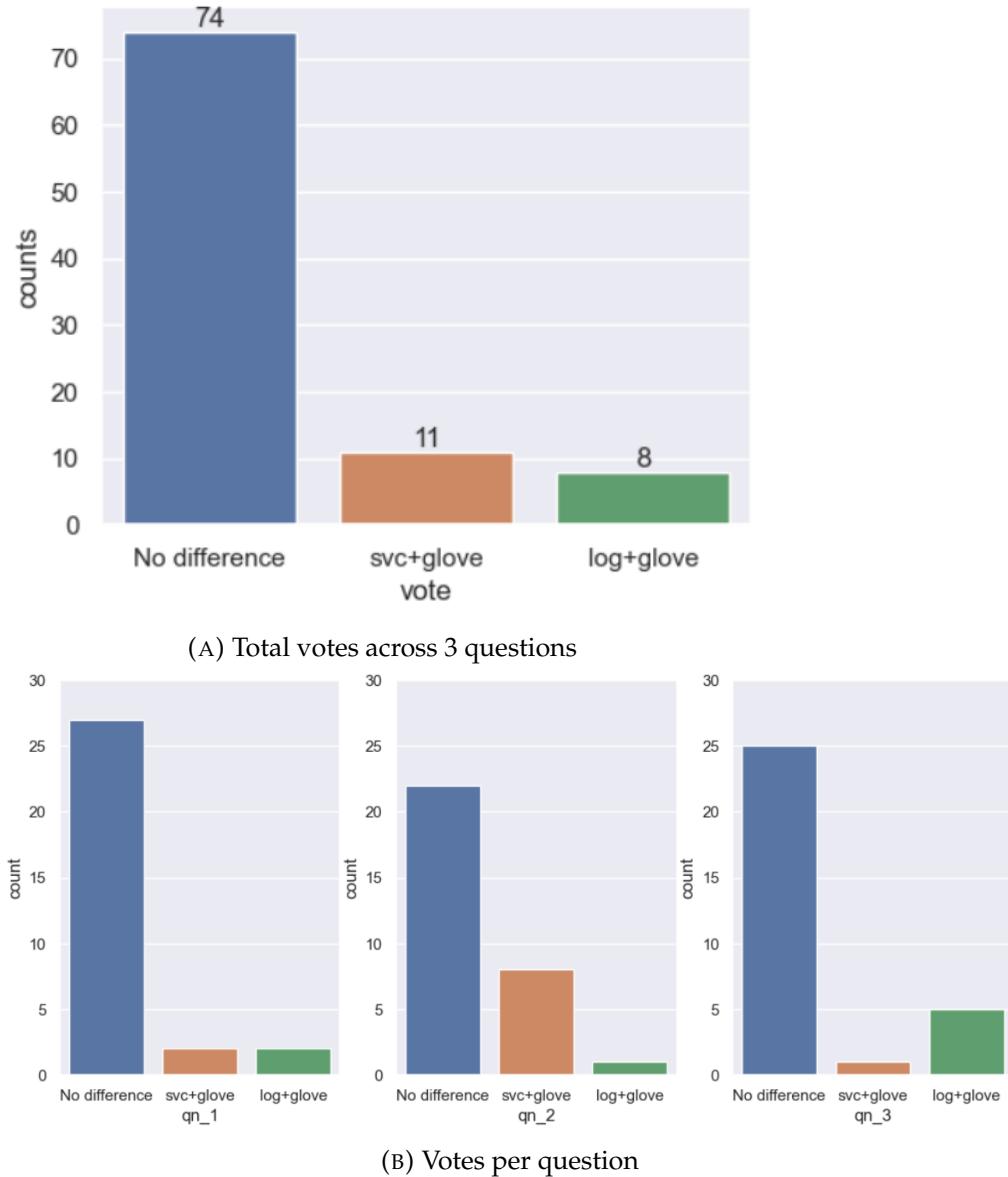


FIGURE 5.4: Testing for which model was more explainable: Respondents' votes to whether SVC + GloVe or Logistic regression + GloVe were more explainable

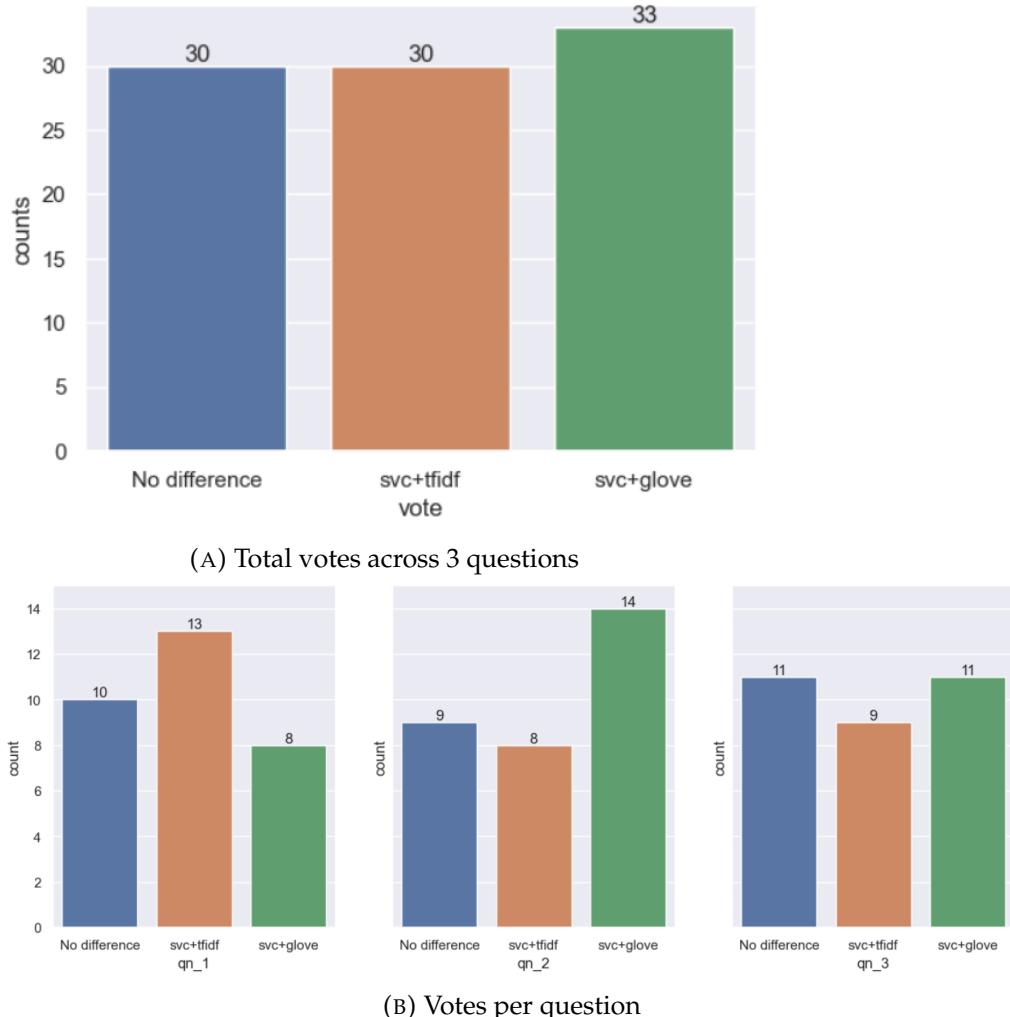


FIGURE 5.5: Testing for which word representation was more explainable: Respondents' votes to whether SVC + TfIDF or SVC + GloVe were more explainable

Bibliography

- 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](https://doi.org/10.1017/9781316761380).
- Bishop, Lea (2023). "A Computer Wrote this Paper: What ChatGPT Means for Education, Research, and Writing". In: *Research, and Writing (January 26, 2023)*.
- Board of Governors of the Federal Reserve System (2022). *Federal Fair Lending Regulations and Statutes Fair Housing Act*. Last Accessed: 2023-2-16. URL: https://www.federalreserve.gov/boarddocs/supmanual/cch/fair_lend_fhact.pdf.
- Bommarito, Michael James and Daniel Martin Katz (2022). "GPT Takes the Bar Exam". In: Available at SSRN. DOI: <https://dx.doi.org/10.2139/ssrn.4314839>.
- 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.
- Council of the European Union (2016a). *Art. 22 General Data Protection Regulation - Automated individual decision making, including profiling*.

- Last Accessed: 2023-2-20. URL: <https://gdpr-info.eu/art-22-gdpr/>.
- Council of the European Union (2016b). *Article 3 General Data Protection Regulation - Territorial Scope*. Last Accessed: 2023-2-20. URL: <https://gdpr-info.eu/art-3-gdpr/>.
- (2016c). *Recital 71 of the General Data Protection Regulation*.
- 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*.
- 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 Shashishekhar 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](https://doi.org/10.1145/3462757.3466145). URL: <https://doi.org/10.1145/3462757.3466145>.
- Greene, Jenna (Dec. 10, 2022). “Will ChatGPT make lawyers obsolete? (Hint: Don’t be afraid)”. In: *Reuters*. URL: <https://www.reuters.com/legal/transactional/will-chatgpt-make-lawyers-obsolete-hint-be-afraid-2022-12-09/> (visited on 02/22/2023).
- 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>.
- Hern, Alex (Oct. 24, 2022). "How Tiktok's algorithm made it a success: 'It pushes the boundaries'". In: *The Guardian*. URL: <https://www.theguardian.com/technology/2022/oct/23/tiktok-rise-algorithm-popularity> (visited on 02/20/2023).
- IBM (2022). *What is natural language processing?* Last Accessed: 2023-02-11. URL: <https://www.ibm.com/sg-en/topics/natural-language-processing>.
- Katz, Daniel Martin et al. (2023). "Natural Language Processing in the Legal Domain". In: Available at SSRN. DOI: <https://dx.doi.org/10.2139/ssrn.4336224>.
- LawGeex (2018). *Comparing the Performance of Artificial Intelligence to Human Lawyers in the Review of Standard Business Contracts*. Last Accesssed: 2023-02-11. URL: <https://images.law.com/contrib/content/uploads/documents/397/5408/lawgeex.pdf>.
- 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>.
- Lohr, Steve (Mar. 19, 2017). "AI Is Doing Legal Work. But It Won't Replace Lawyers, Yet." In: *The New York Times*. URL: <https://www.nytimes.com/2017/03/19/technology/lawyers-artificial-intelligence.html> (visited on 02/22/2023).
- 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/>.
- 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>.
- 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>.
- Smith, Ben (Dec. 5, 2021). "How TikTok Reads Your Mind". In: *The New York Times*. URL: <https://www.nytimes.com/2021/12/05/business/media/tiktok-algorithm.html> (visited on 02/20/2023).
- Statista (2022). *Legal tech market revenue worldwide from 2021 to 2027*. Last Accessed: 2023-02-11. URL: <https://www.statista.com/statistics/1155852/legal-tech-market-revenue-worldwide/>.
- Wagner, Isabel (2022). "Privacy Policies Across the Ages: Content and Readability of Privacy Policies 1996–2021". In: *arXiv preprint arXiv:2201.08739*.

- 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 explainability 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



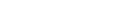
4



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. T *

hi
n
k
th
at
u
si
n
g
th
e
m
o
d
el
is
a
n
ef
fe
ct
iv
e
m
et
h
o
d
of
i
d
e
nt
if
yi
n
g
vi
ol
at
io
n
s
of
th
e
P
D
P
A
?

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



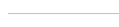
2



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:

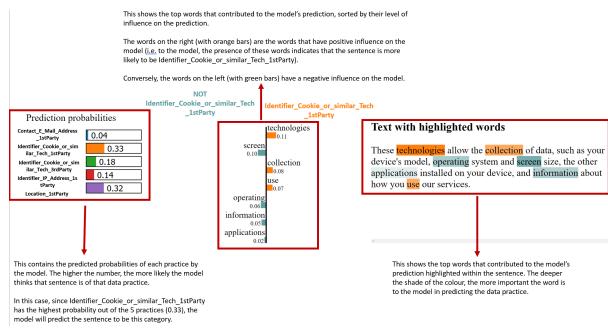
The model predicted the practice wrongly.

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. D *
 i
 d
 y
 o
 u
 fi
 n
 d
 t
 h
 e
 vi
 s
 u
 a
 li
 s
 a
 ti
 o
 n
 e
 a
 s
 y
 t
 o
 i
 n
 t
 e
 r
 p
 r
 e
 t
 ?

Mark
only
one
oval.

Very difficult

1
 2
 3
 4
 5

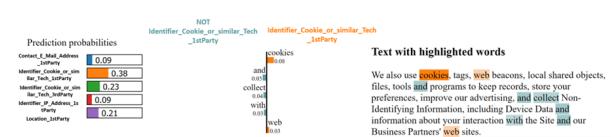
Very easy

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".

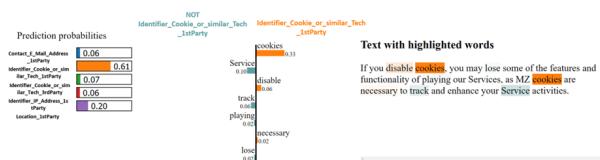
Once again, there is no "correct" option. Simply choose whichever option that you agree with the most.

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*
 hi
 c
 h
 e
 x
 pl
 a
 n
 at
 io
 n
 di
 d
 y
 o
 u
 fi
 n
 d
 e
 a
 si
 er
 to
 in
 te
 rp
 re
 t?

Mark
only
one
oval.

4

.
 1
 (.
 i
)

4

.
 1
 (.
 i
 i
)

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

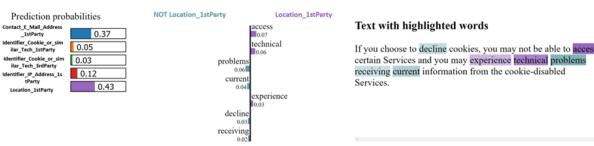
Visualisation 4.2(i)

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.

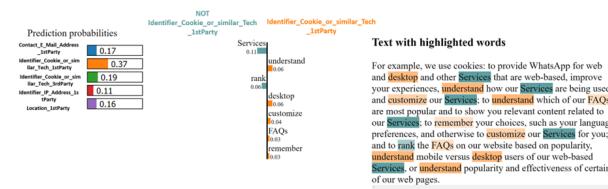
- 4.2(i)
- 4.2(ii)
- No difference

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

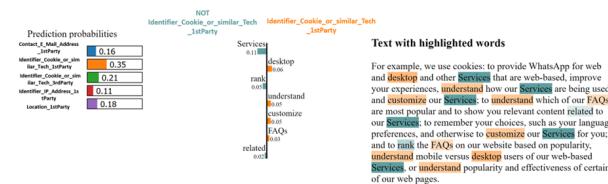


Visualisation 4.3(ii)

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.

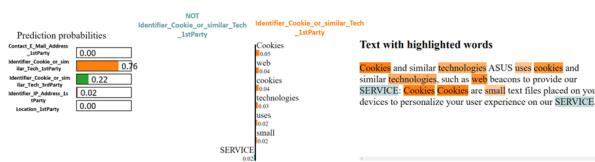
Choose "no difference" if the visualisations seem the same to you.

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

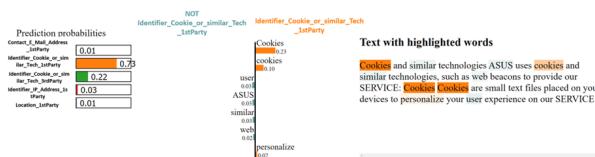


Visualisation 5.1(ii)

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

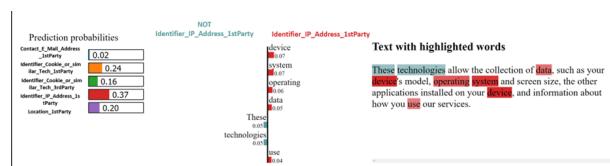
Visualisation 5.2(i)

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.

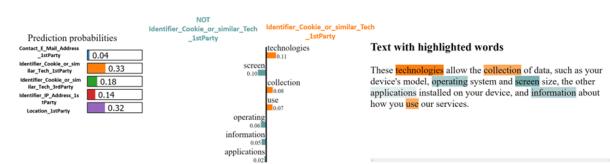


Visualisation 5.2(ii)

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

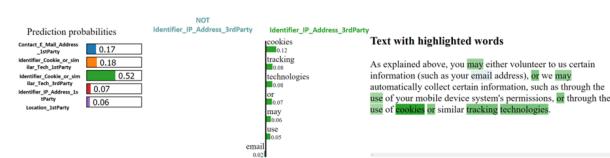
Visualisation 5.3(i)

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.

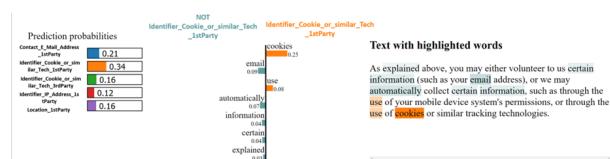


Visualisation 5.3(ii)

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. T *

hi
n
k
th
at
u
si
n
g
th
e
m
o
d
el
is
a
n
ef
fe
ct
iv
e
m
et
h
o
d
of
i
d
e
nt
if
yi
n
g
vi
ol
at
io
n
s
of
th
e
P
D
P
A
?

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

This content is neither created nor endorsed by Google.

Google Forms

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](#)

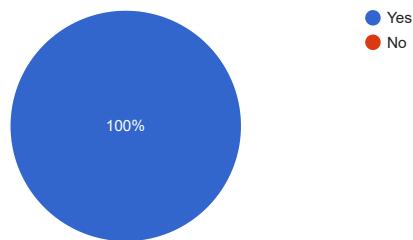
If you are a NUS / Yale-NUS student, are you 18 years old and above?

[Copy](#)

OR

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

31 responses

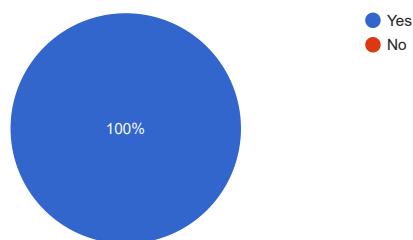


Participant Information Sheet

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

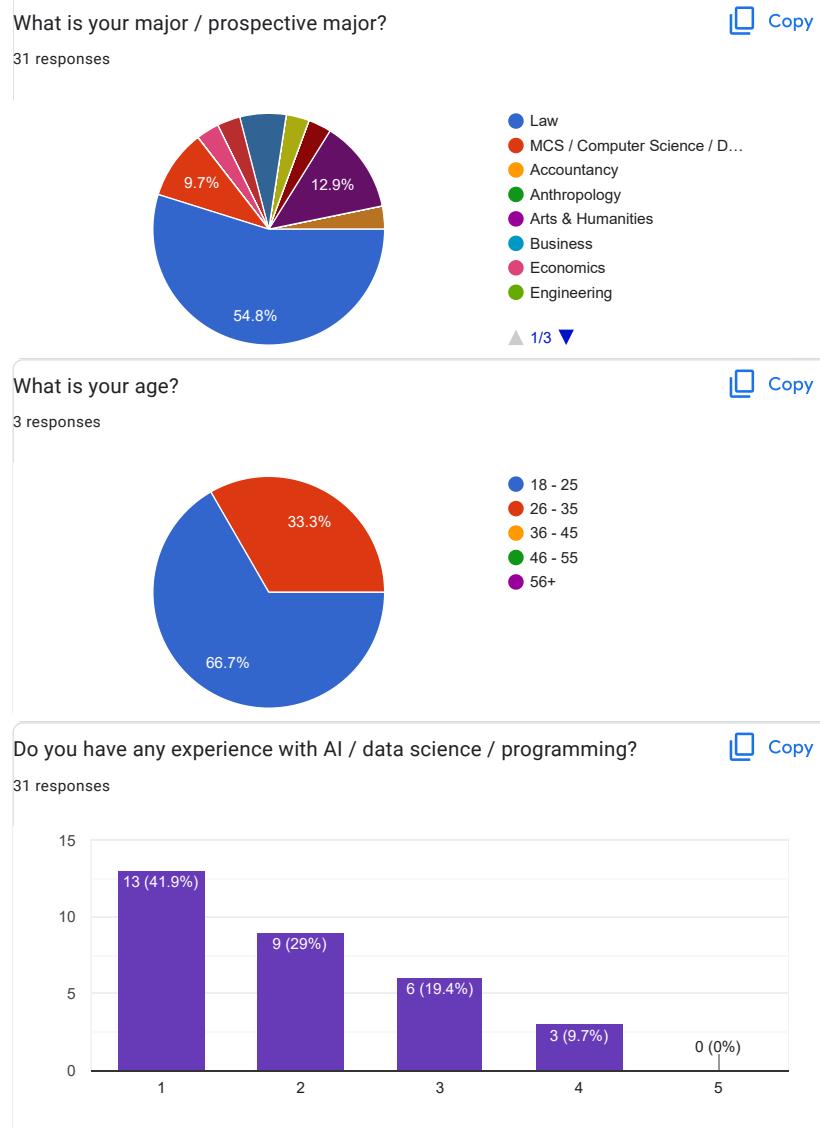
[Copy](#)

31 responses

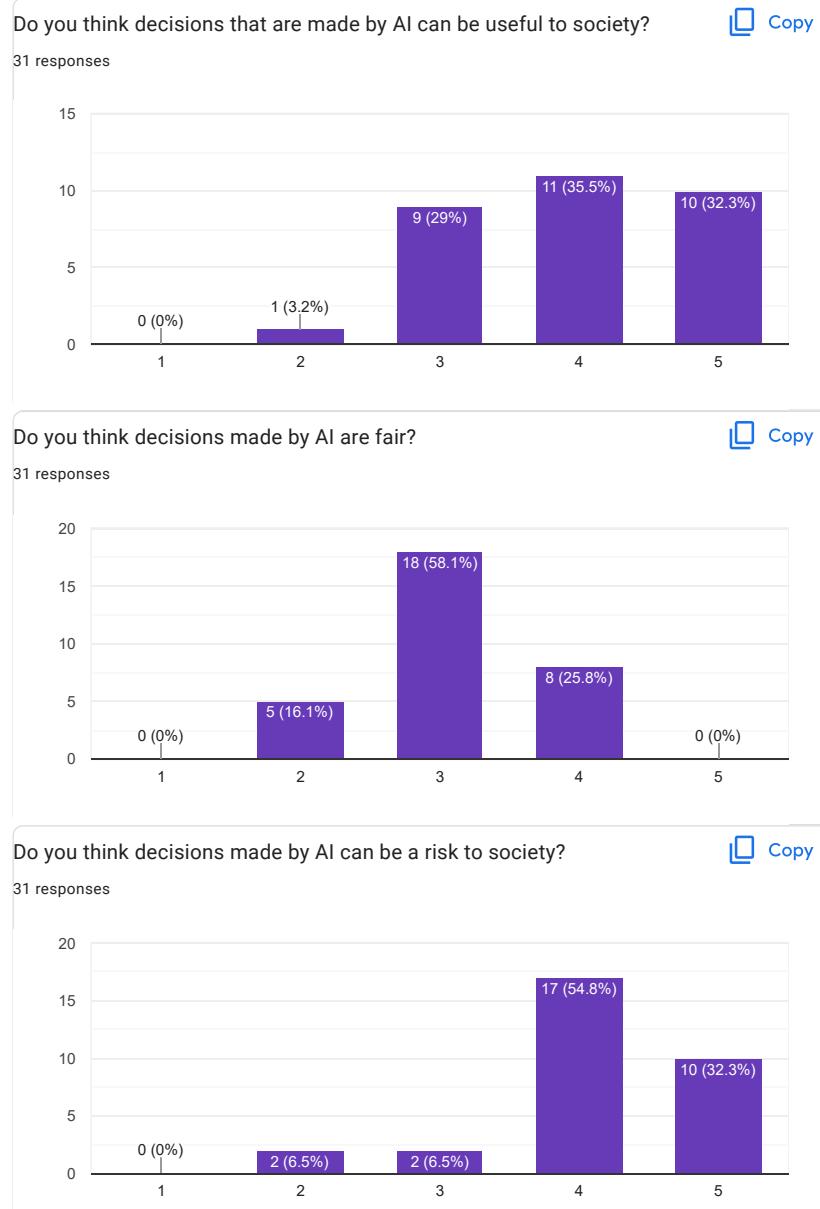


Part 1









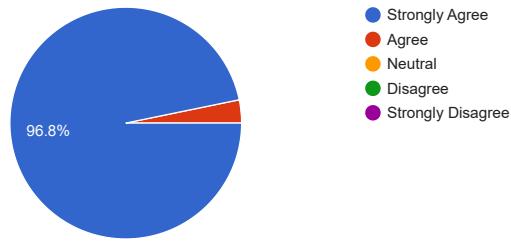
Part 2



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

 Copy

31 responses



- 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.

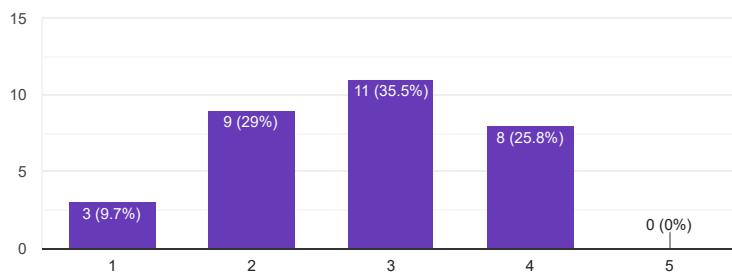
If the pre-drafted data privacy policy actually does not state that cookies are being used but your app uses cookies, you could face a fine of up to \$10,000 in breach of the PDPA as you would have failed to notify your users.

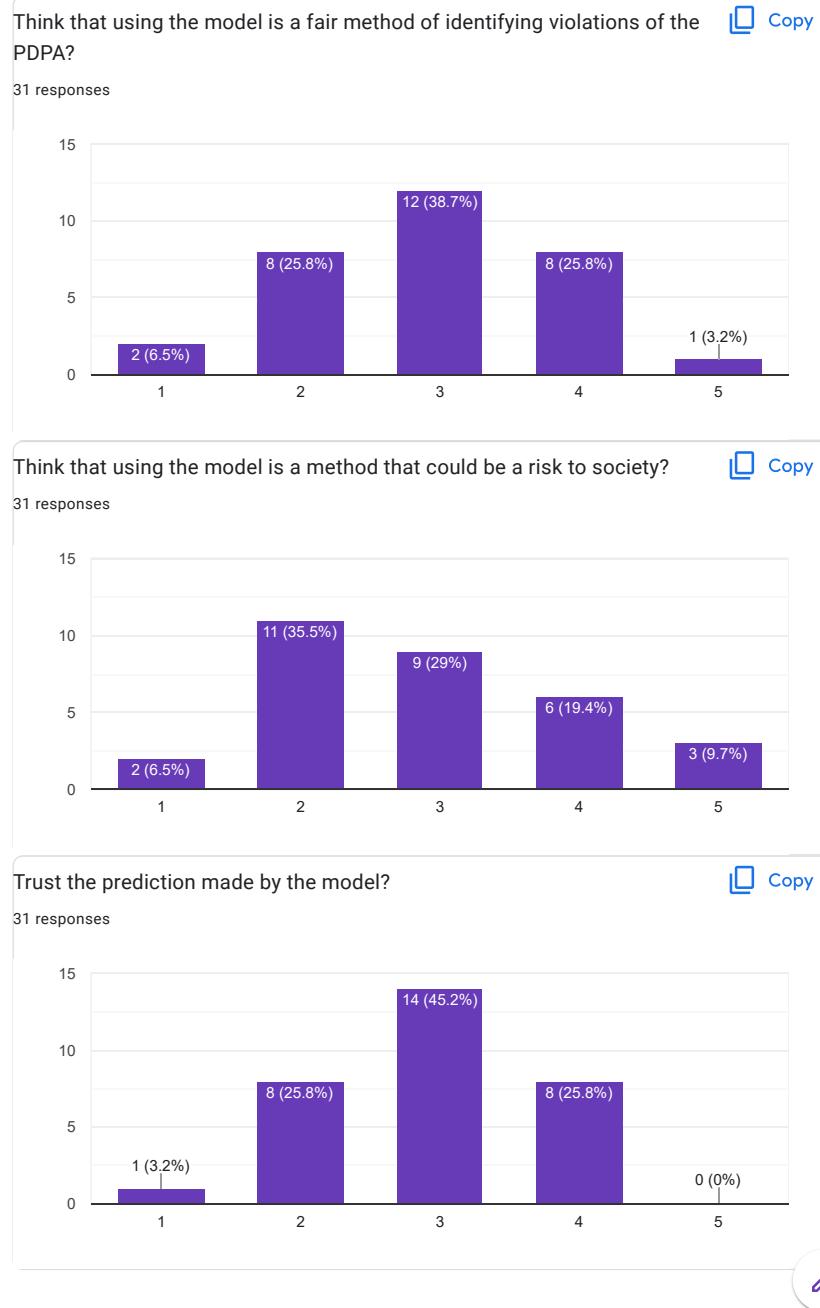
How far do you, as the app developer:

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

 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.

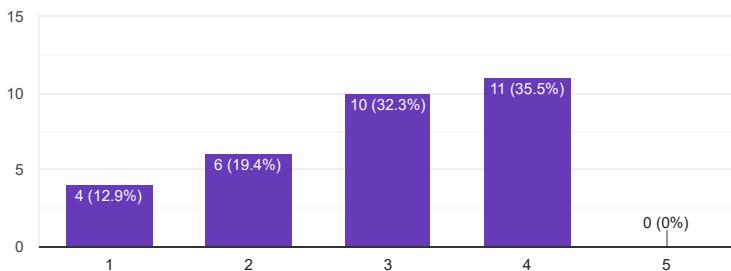
To increase the efficiency of the PDPC, your team is considering whether to adopt the abovementioned model to automate the analysis of data privacy policies. If this new method of analysis is adopted, the PDPC would rely entirely on the model's predictions to confirm whether app developers have breached 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?

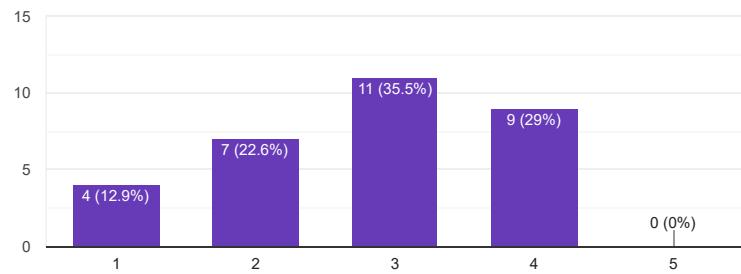
 Copy

31 responses



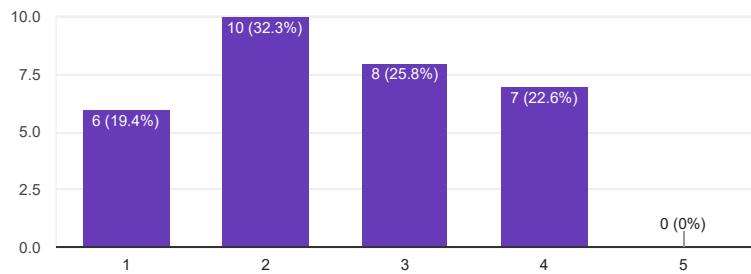
Think that using the model is a fair method of identifying violations of the PDPA? [Copy](#)

31 responses



Think that using the model is a method that could be a risk to society? [Copy](#)

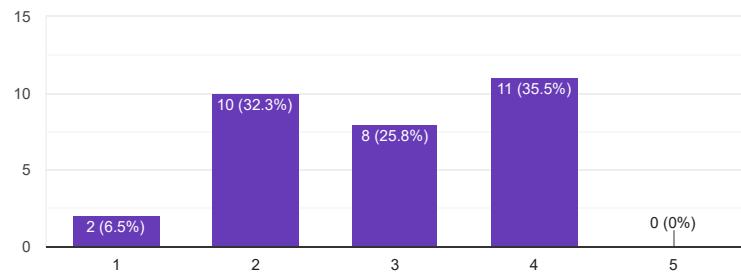
31 responses



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.

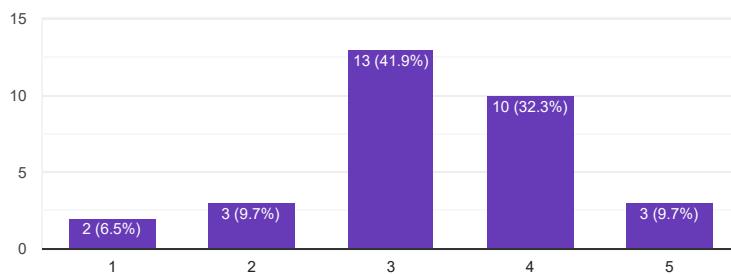
If the PDPC decides that the developer has indeed violated the PDPA, you could claim compensation from the app developer of up to \$10,000.

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?

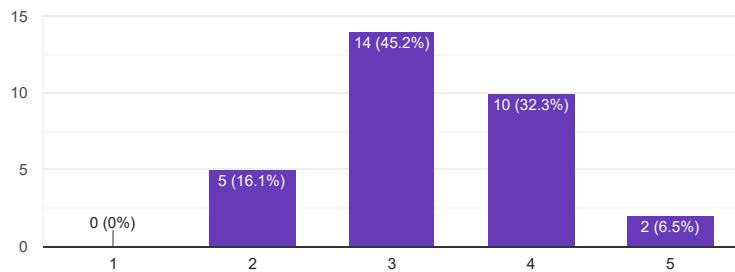
 Copy

31 responses



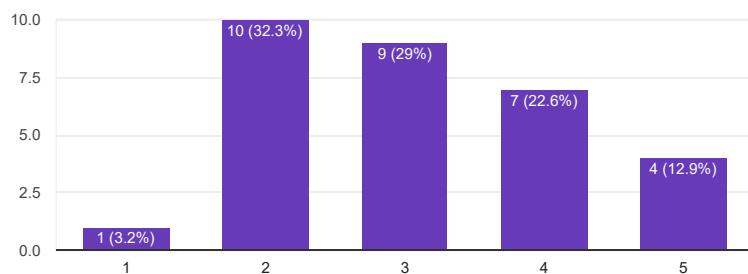
Think that using the model is a fair method of identifying violations of the PDPA? [Copy](#)

31 responses



Think that using the model is a method that could be a risk to society? [Copy](#)

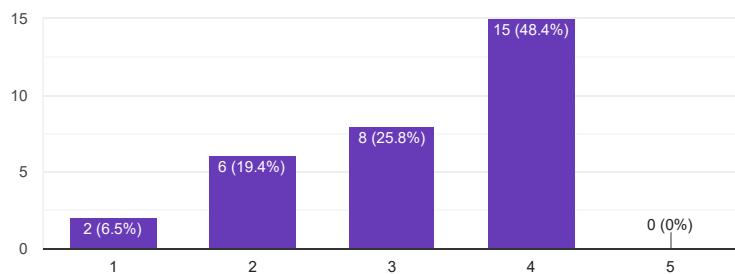
31 responses



Trust the prediction made by the model?

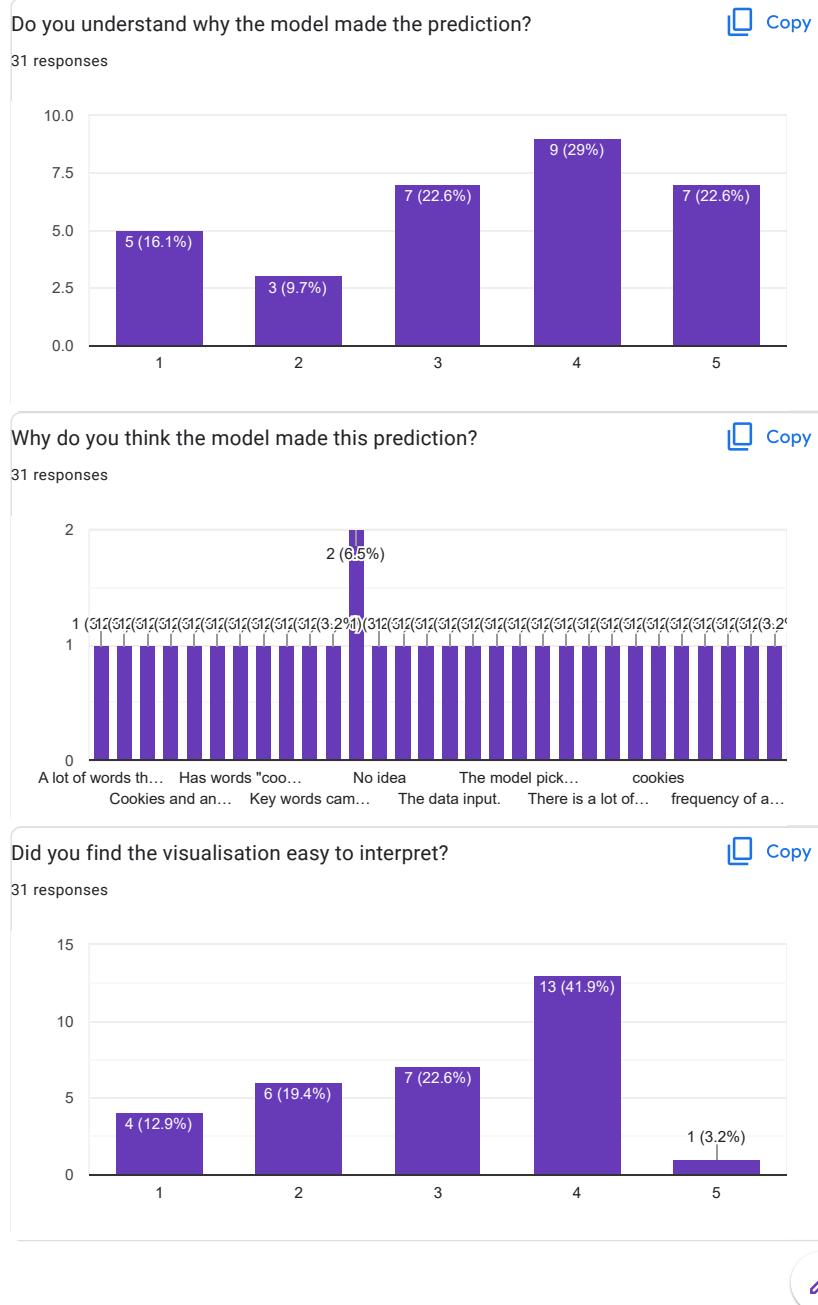
[Copy](#)

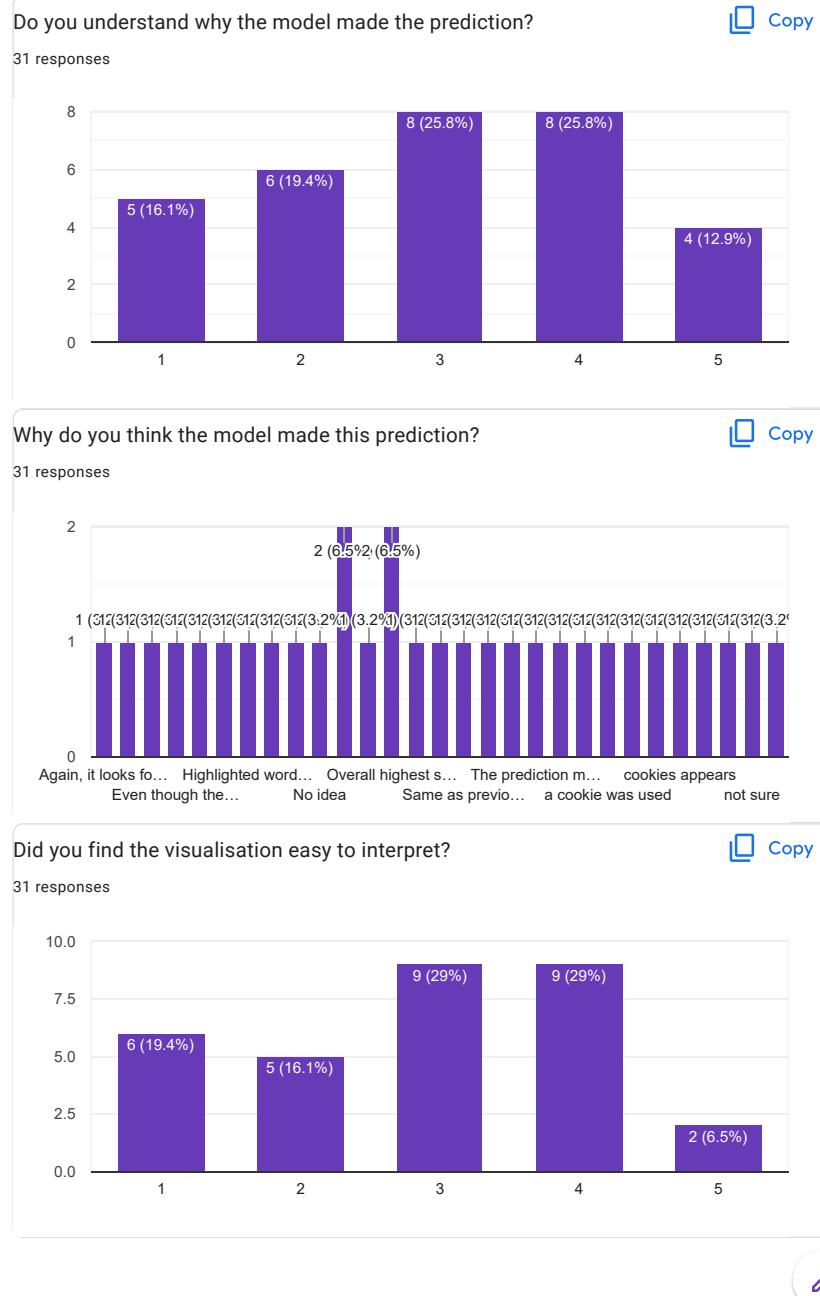
31 responses



Part 3





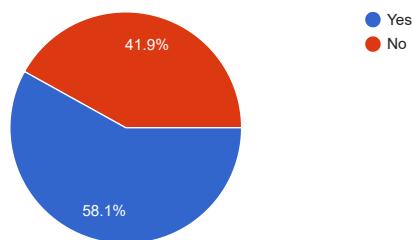


Based on your current understanding, do you think that the sentence below would be predicted to be "Identifier_Cookie_or_Similar_Tech_1stParty"?

 Copy

"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."

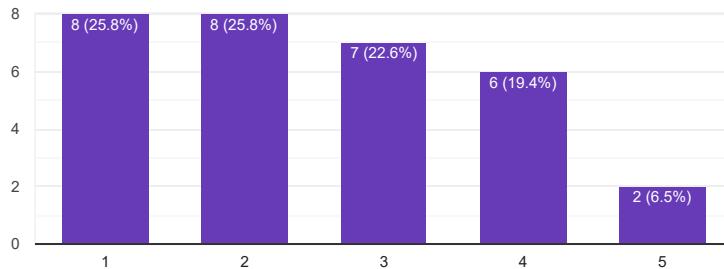
31 responses

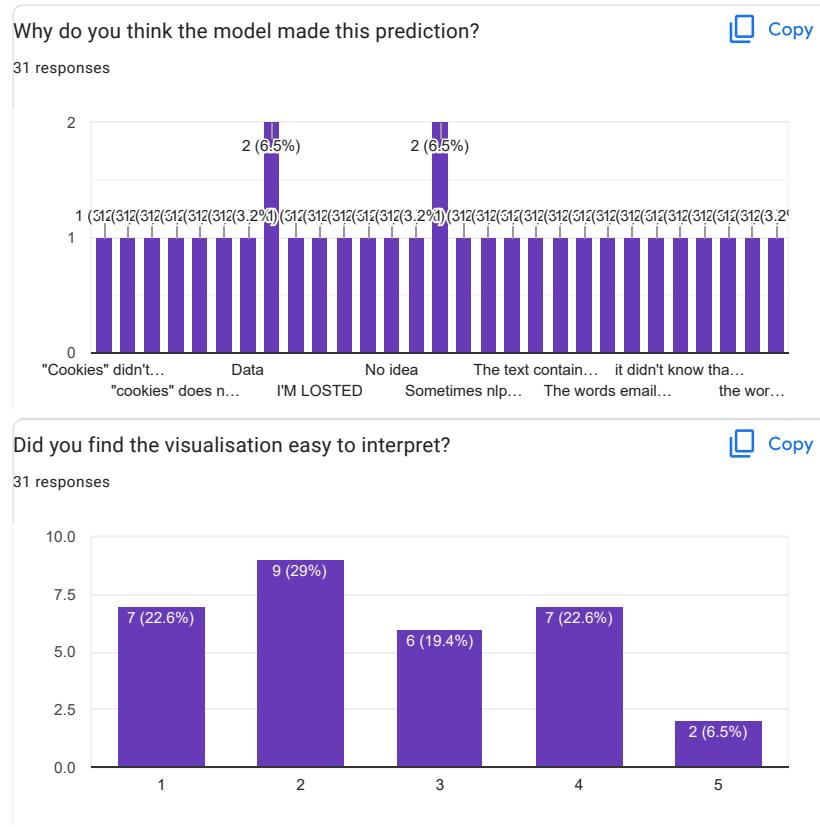


Do you understand why the model made the prediction?

 Copy

31 responses



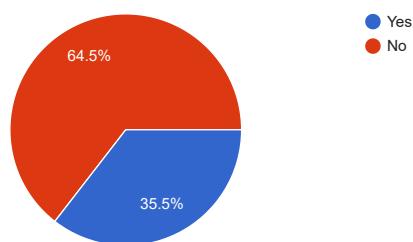


Based on your current understanding, do you think the sentence below would be predicted to be in "Identifier_Cookie_or_Similar_Tech_1stParty?"

 Copy

"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."

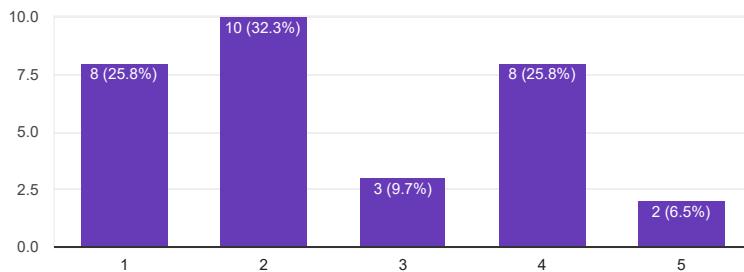
31 responses

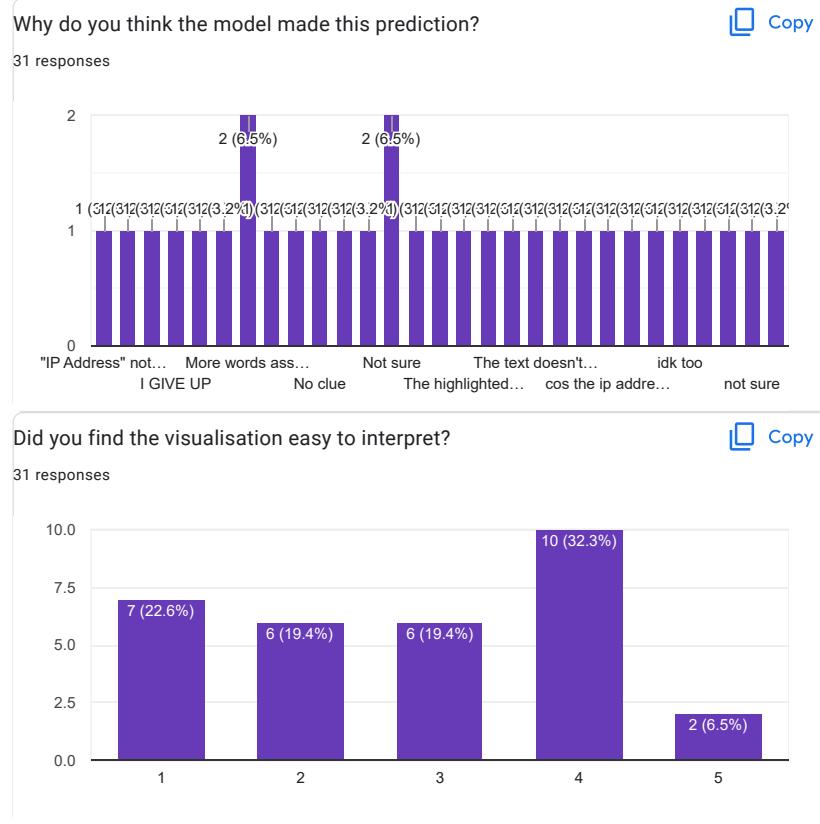


Do you understand why the model made the prediction?

 Copy

31 responses



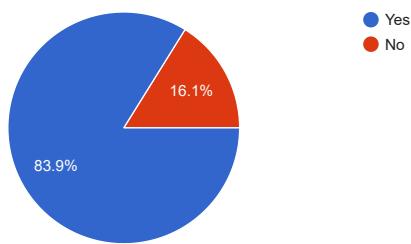


Based on your current understanding, do you think the sentence below would be predicted to be in "Identifier_Cookie_or_Similar_Tech_1stParty?"

 Copy

"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."

31 responses

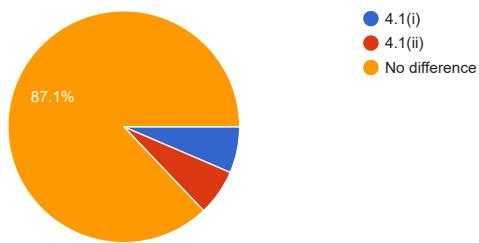


Part 4

Which explanation did you find easier to interpret?

 Copy

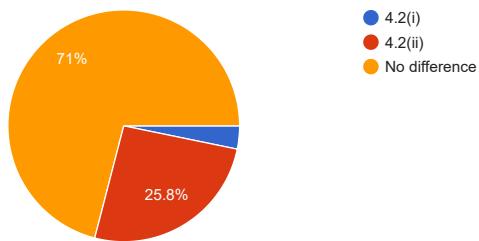
31 responses

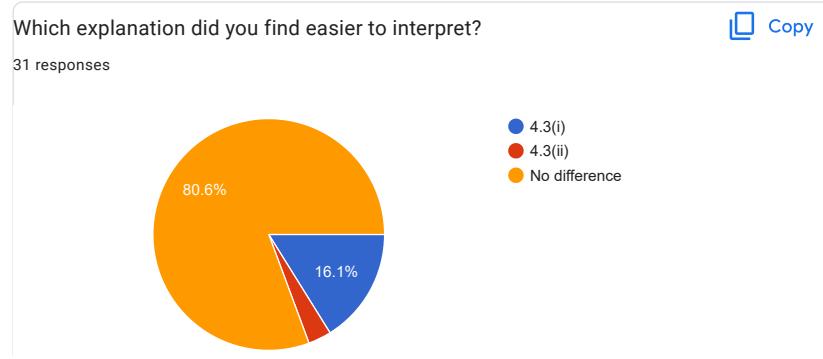


Which explanation did you find easier to interpret?

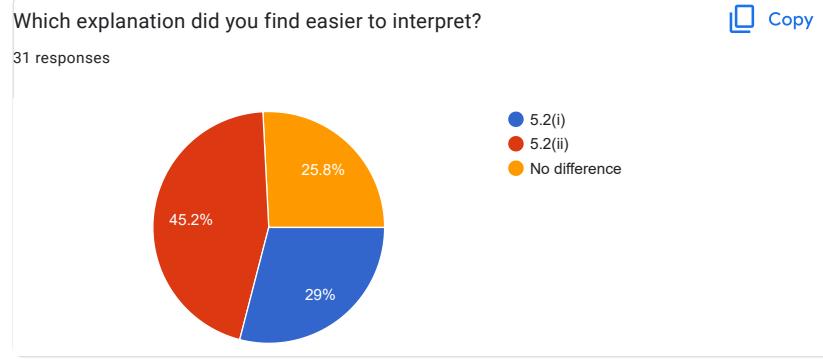
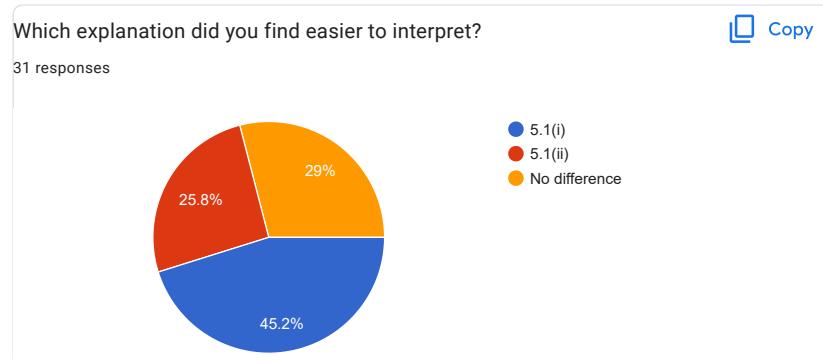
 Copy

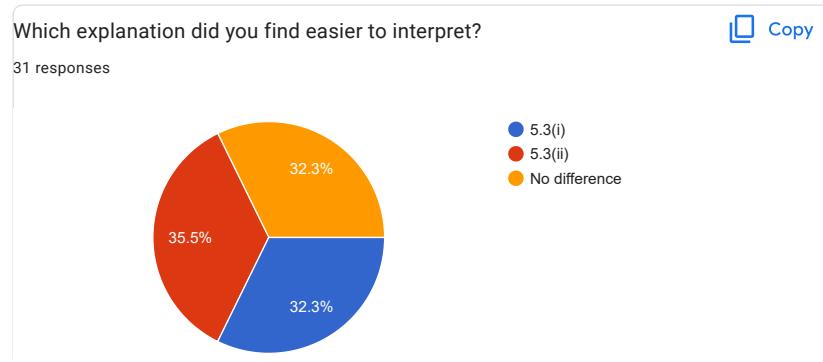
31 responses





Part 5





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.

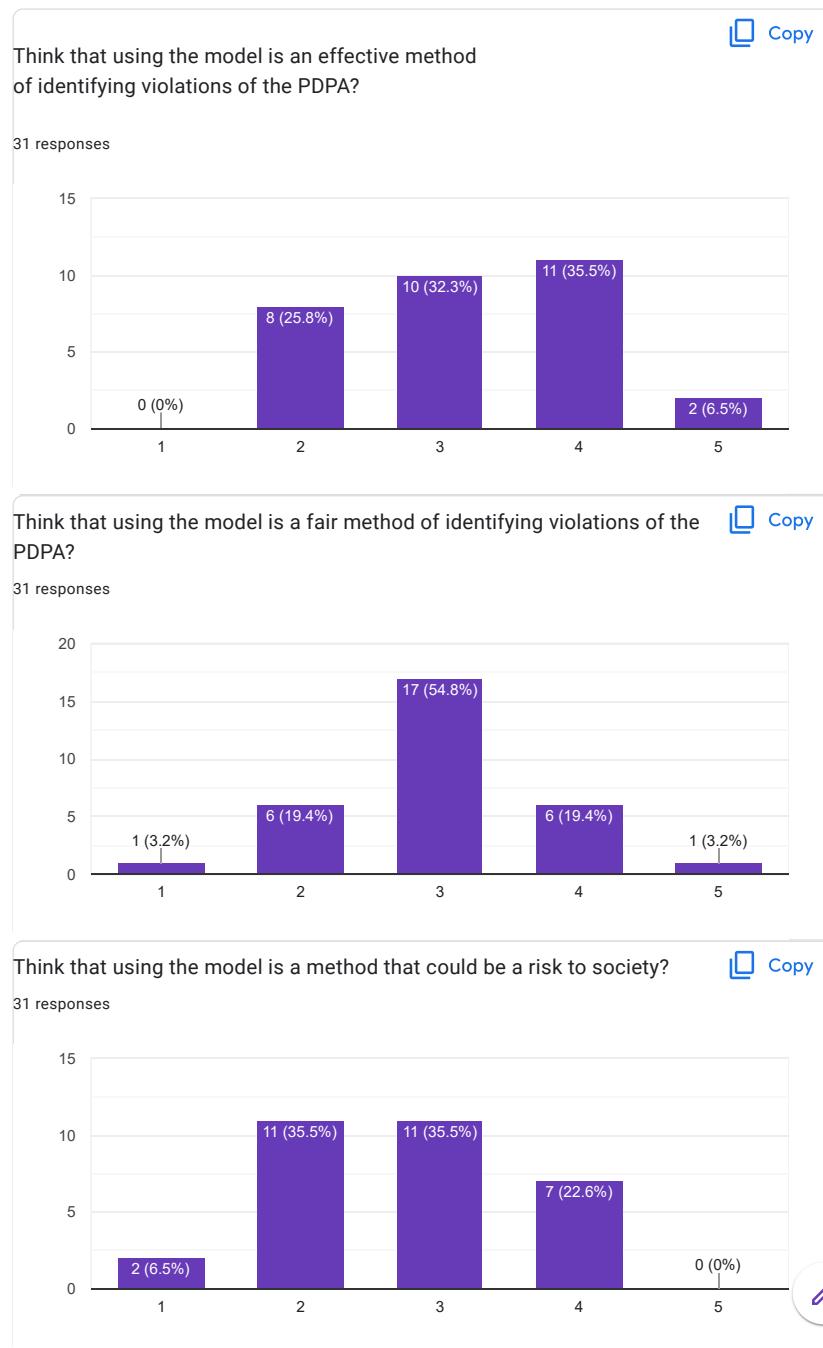
Since you have no knowledge of the PDPA, you use the abovementioned model to analyse a pre-drafted data privacy policy that you found online. The model informs you that there is a sentence which states that cookies are being used.

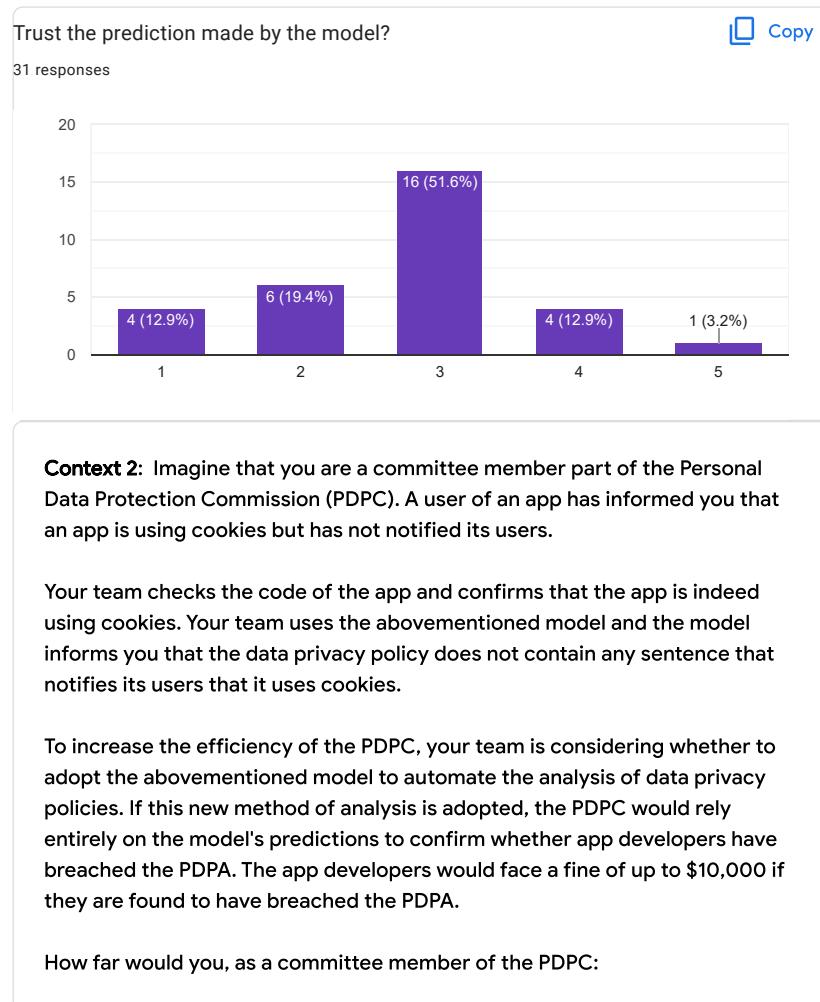
You are deciding whether to rely entirely on the model's prediction, or pay costly legal fees to confirm with your friend who is a lawyer.

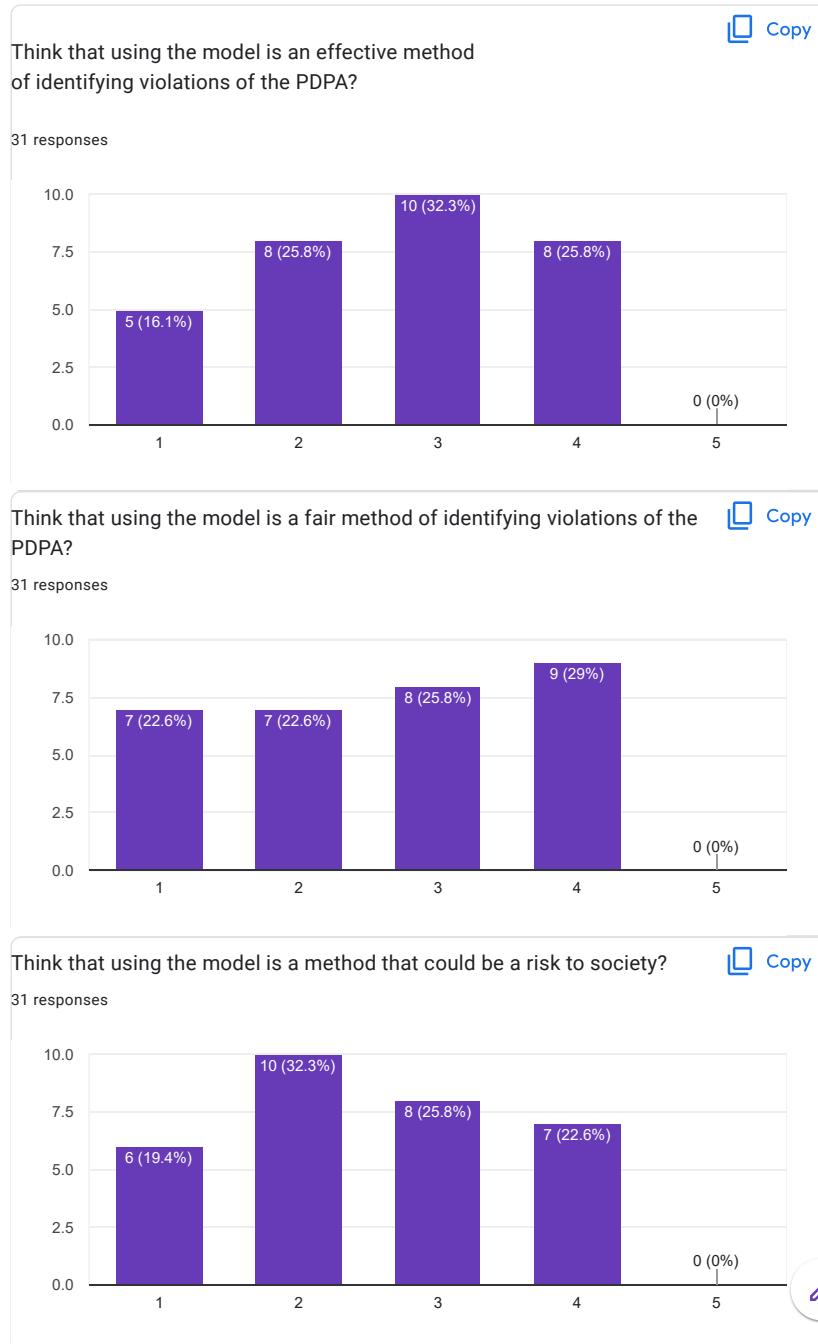
If the pre-drafted data privacy policy actually does not state that cookies are being used but your app uses cookies, you could face a fine of up to \$10,000 in breach of the PDPA as you would have failed to notify your users.

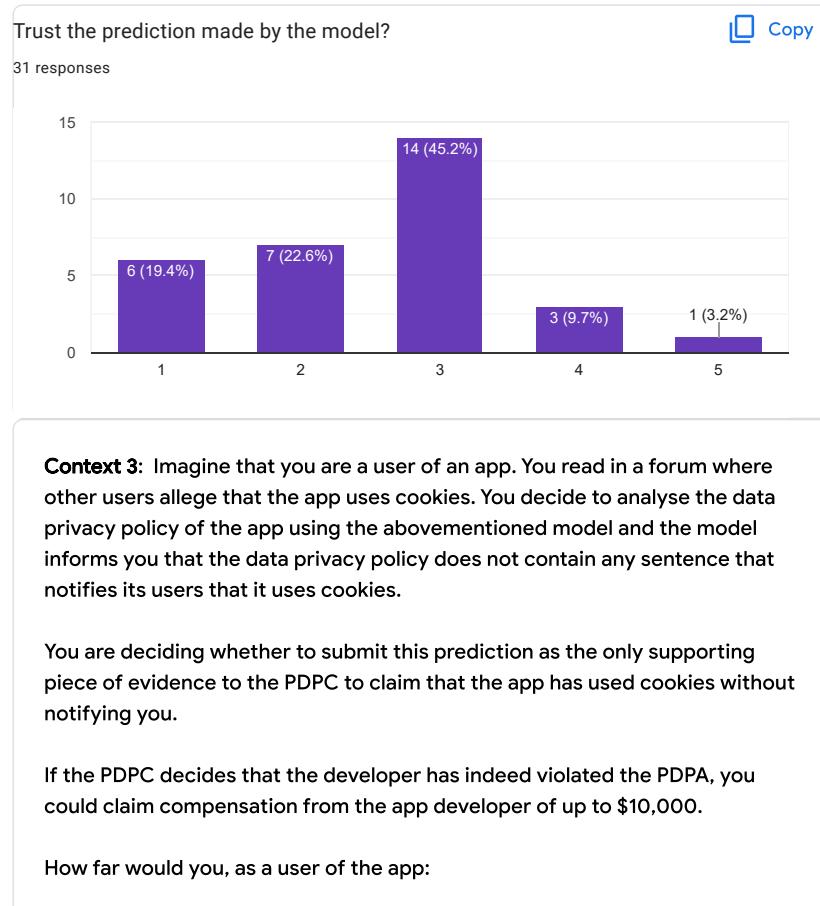
How far do you, as the app developer:

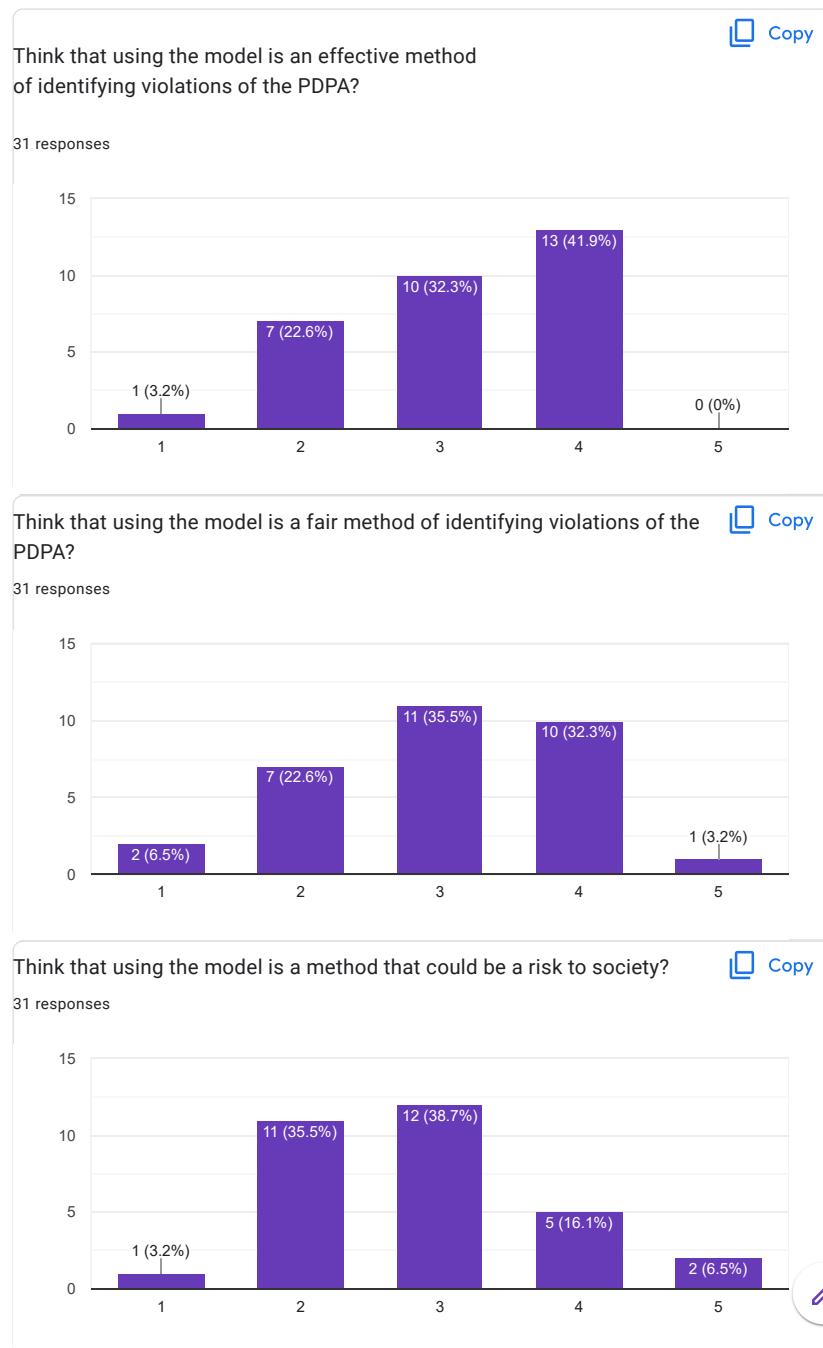


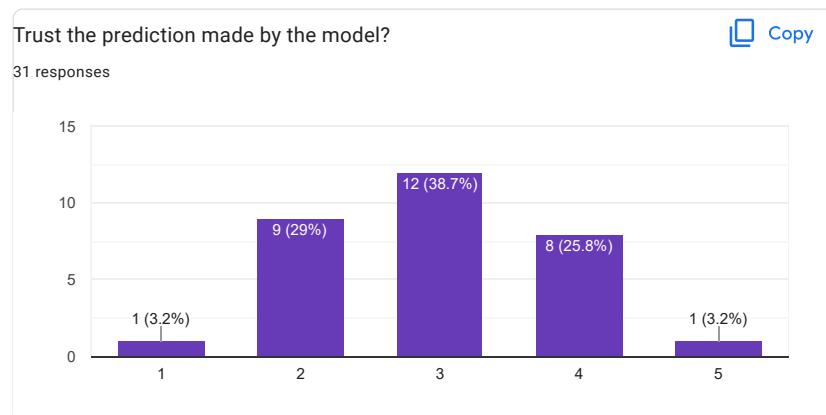












This content is neither created nor endorsed by Google. [Report Abuse](#) - [Terms of Service](#) - [Privacy Policy](#)

Google Forms



