RE: PRIVACY;
INVESTIGATING THE
ACCESSIBILITY OF ONLINE LEGAL
DOCUMENTS IN RELATION TO THE
PRIVACY PARADOX

BY
ELIZABETH IVY CALVERT

A 90-POINT THESIS SUBMITTED TO THE VICTORIA
UNIVERSITY OF WELLINGTON IN PARTIAL
FULFILMENT OF THE REQUIREMENTS FOR THE
DEGREE OF **MASTER OF DESIGN INNOVATION**

VICTORIA UNIVERSITY OF WELLINGTON
2018

ADDITIONAL THESIS CONTENT

Dear Reader,

This thesis includes a code output. This code is available at:
https://github.com/ivycalvert/Re-Privacy_Tool

It is recommended that you consider visiting this website and exploring the creative output as you read through this thesis. While overall this thesis can be considered a complete body of work without the supplement of the code, it is suggested that you also explore the creative output yourself.

If that URL is inaccessible, please request the digital files submitted alongside this thesis. They contain a full copy of the code, and instructions on how to run the code from your computer.

FOREWORD

After growing up in rural New Zealand, I was first properly exposed to widespread internet integration into society during tertiary education. Coming from a childhood that largely supported personal privacy, this thesis was inspired by the surprise felt as I first began to realise how openly many internet users disclose personal information on the internet.

This led to research addressing “the biggest lie on the internet”—the societal expectation that very few internet users read the Terms of Service and Privacy Policy documents attached to the websites they interact with online.

Coming from a design focused background, addressing issues that touch on multiple fields (law, data science, sociology, and others) allows for an informed human-computer interaction approach to addressing this issue. The combination of practical and theoretical practice that design offers allows for a critical evaluation of the issue, the ability to build the functionality of a possible solution, and the user interface that could inspire user adoption of the solution. This experience, paired with my personal motivation and investment in the subject, have encouraged the development of this thesis and design output.

ACKNOWLEDGEMENTS

Thank you to those who tolerated me through the duration of this thesis.

TABLE OF CONTENTS

<i>i</i>	Additional Thesis Content
<i>iii</i>	Foreword
1	Abstract
5	Key Concepts
	<i>Privacy Paradox</i>
	<i>Personally Identifiable Information</i>
	<i>Terms of Service</i>
	<i>Privacy Policy</i>
	<i>Machine Learning</i>
13	Introduction
	<i>Thesis Structure</i>
19	Context
	<i>Precedent Analysis</i>
39	Methodology
	<i>Content Analysis</i>
	<i>Case Studies</i>
	<i>Questionnaire</i>
	<i>Configuration</i>
	<i>Prototyping</i>
45	Regulation and Legislation
	<i>Case Studies</i>
	<i>Socio-technical Aspects</i>
57	Questionnaire
	<i>Previous Surveys</i>
	<i>Participants</i>
	<i>Distribution</i>
	<i>Limitations</i>
	<i>Results</i>
71	Design Ideation
	<i>Process</i>
	<i>Iterations</i>
	<i>Scope</i>
89	Re: Privacy (Proof of Concept Tool)
	<i>Browser Extension</i>
	<i>ML Model</i>
	<i>Integration</i>

99	Discussion
105	Conclusion
111	Bibliography
121	List of Figures
127	Appendix
	<i>Ethics Application</i>
	<i>Questionnaire questions</i>

ABSTRACT

Website Terms of Service and Privacy Policy documents are the delivery methods employed by online services to inform internet users of how information is treated on their platform. When users fail to read or understand these documents, many issues can arise. These issues take shape in the form of unwillingness to use services, misunderstandings of how online data is treated, and/or user concerns for personal privacy. Currently, the leading factors influencing user motivation to read these documents include (but are not limited to) document length, complicated language, and time required to read.

To encourage users to engage with these documents, this research investigates the delivery methods service providers use to present these legal documents online. During the course of this investigation, the proposal of a new method of presenting website Privacy Policies to users is explored. This tool, 'Re: Privacy', was developed with the aim of increasing user awareness of online data treatment, whilst minimising the factors that dissuade users from reading the official document.

Published alongside Re: Privacy is a comprehensive analysis of website legal documents and user understanding of these documents. This analysis also provides an investigation into user understanding of current website legal documents, and how user understanding can be improved.

KEY CONCEPTS

As with any research, this thesis contains a considerable amount of field-specific terminology. For ease of reading, the words and terms of most importance will be explained and defined here. These terms will also be discussed in further depth within the context of the thesis as they become relevant.

PRIVACY PARADOX

The Privacy Paradox is a phenomenon that refers to when an internet user's online privacy concerns are not reflected in their actions online. Primarily, this is when a user expresses that they experience concerns for their privacy online, but take little to no actions to protect the personal information they self-disclose online. However, the privacy paradox is not limited solely to self-disclosure, but also includes when users' experience concerns about the personal information other users disclose about them online without their consent.

PERSONALLY IDENTIFIABLE INFORMATION (PII)

Personally Identifiable Information is any information that can be used to identify, contact, or locate a single person. This can include but is not limited to: email address, location, social security number, name, age, race, sex, weight, height, relationship status, education level, politics, buying habits, household health concerns (Roderick, 2014). Any information that can be used to specifically identify any single person is included in this list.

TERMS OF SERVICE (ToS)

Terms of Service are legal documents which are also commonly known as Terms of Use or Terms and Conditions. ToS outline the rights of each party which people must agree to in order to use and/or continue using a digital service. They are a form of digital contracting that allows services to be offered through digital mediums while still providing a legal outline that protects users' and service providers' rights.

PRIVACY POLICY (PP)

A Privacy Policy is an agreement where the service provider must specify what personal information is collected from the users and how data is used/protected by the service provider. These documents are required by law to be provided by services as this information is considered personally identifiable information (defined above). Carefully constructed privacy policy statements are used to inspire trust between internet users and service providers and help users make informed decisions about which companies to engage with.

MACHINE LEARNING (ML)

Machine learning is a subfield of computer science that focuses on the construction of algorithms to allow the program to learn and make predictions based off specified data. This allows the program to run and complete a task without explicit programming (Samuel, 1959). ML technology is often used for automation purposes as the trained models are introduced to new data, they are able to adapt.

"...the Party imposed itself more successfully on people incapable of understanding it. They could be made to accept the more flagrant violations of reality, because they never fully grasped the enormity of what was demanded of them, and were not sufficiently interested in public events to notice what was happening. By lack of understanding they remained sane. They simply swallowed everything, and what they swallowed did them no harm, because it left no residue behind, just as a grain of corn will pass undigested through the body of a bird."

(Orwell, 1949, p.180)

INTRODUCTION

The term Privacy Paradox refers to a social phenomenon which has grown in prominence since the term's first use in 2006 (Barnes, 2006). Originally the term was used to explain young internet users oversharing to online predators, however this has since been adapted to explain the ambiguous boundary between what is private and what is public on the internet. This refined definition places particular attention to privacy concerns expressed by internet users and their failure to act on these concerns. Currently, this phenomenon persists due to the popularity of internet services that require internet users to upload Personally Identifiable Information (PII) in order to interact or gain access to those services.

Personally Identifiable Information (PII) refers to information which can be used to identify, contact, or locate a single person. Theodore Roszak (1986, p.186) stated that the collection of PII "reduce[s] people to statistical skeletons for rapid assessment... No ambiguities, no subtleties, no complexities. The information that data banks hold is life stripped down to the bare necessities required for a quick commercial or legal decision". PII is often disclosed online by internet users in order to gain access to or personalise online services.

To understand the privacy paradox it is important to understand the multitude of factors that have resulted in the phenomenon and how the numerous services offered online has encouraged its development. Numerous types of online services encourage users to self-disclose PII online. "Every time you use your smartphone, surf the web, click on an advertisement, post a comment, picture or video, track your sleep, moods or runs, the digital record of 'you' will become more detailed" (Lanzing, 2016, p.85). This in itself, is not inherently bad, but rather offers a range of positives and negatives that affect internet users. Users avidly engage with online services as the internet offers new ways to participate and maintain relationships, enable remote commerce, access a wealth of knowledge, provide entertainment, and much more. Internet users are becoming aware that the information they disclose both exposes them to potential loss of privacy, and also contributes to social capital online (Hallam & Zanella, 2016). This awareness has given rise to

privacy concerns surrounding the protection of personal information online; in turn contributing to the development of the privacy paradox phenomenon.

With limited exploration into a solution to eradicate the privacy paradox, this research proposes a shift of focus away from the effects and response to the privacy paradox. This research will instead investigate how the legal policies of service practices can be adjusted to help make these more accessible for internet users. With this overarching topic, this research will focus on the question:

- *How can media design be used to inform internet users of the impact of sharing personal information online?*

RESEARCH AIMS AND OBJECTIVES

While investigating the aforementioned research question, the aims and objectives were as follows:

Investigate the connection between the privacy paradox and online website terms of service and privacy policy documents.

- Situate the research within the current bodies of literature on the privacy paradox and website legal documents. Focusing on why these research topics are connected;
- Define the areas where current terms of service and privacy policies provide an area for the privacy paradox to exist;
- Describe the issues that exist around current website terms of service delivery methods and engagement issues and the current methods of resolving these.

Create a browser extension which focuses on summarising terms of service and privacy policy documents in plain english for internet users.

- Encourage internet users to read website terms of service and better educate them on the treatment of personal data online;
- Increase the accessibility and readability of website terms of service.

In summary, this research aims to provide an output that creates a summary of website legal documents in plain english to make it easier for internet users to understand the content.

THESIS STRUCTURE

Due to the two main aims of the thesis providing two clear parts, these will be used as a division in the thesis. This places an emphasis on correct identification and investigation into a problem, which informs complete, dedicated production of a design solution.

The first section of the thesis will focus on situating the research and investigating the research fields. This “Context” section will also focus on data collection and assumption checking prior to creating the final design

output (an important step to ensure the validity of the research output and final conclusions).

The second section of the thesis focuses on the iterative prototyping and development of the final design output. It is important to completely establish the context of the thesis before developing the final design output to ensure the research can inform the design during the development stage. Although, as stated, it is important for the final output to be informed by the final conclusions of the contextual research. It is also important to begin the ideation and conceptual generation of the design output alongside the contextual research. This ensures the concept generation is continuous and enables both the contextual research and the design output to be developed as a cohesive whole.

CONTEXT

The World Wide Web was first developed by Tim Berners-Lee in 1989 (Hafner & Lyon, 2001) but it wasn't until after it was released free to the public in 1993 that it experienced growth on a large scale. The work of Berners-Lee enabled easier navigation of the internet as computer scientists developed point-and-click programs (Hafner & Lyon, 2001), with browser programs such as 'Mosaic' popularising the technology by making "the software easy to use and appealing to 'non-geeks'" ("The National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign," n.d., para. 1).

With the early 1990s, marking the beginning of the internet era, most children born from this time onwards have lived the majority or all of their lives with access to the internet; marking the first digital native generation. Within the first few years of the internet, online services such as search engines were a main development focus (Hafner & Lyon, 2001), however after the success of large companies such as Google (founded in 1998), it wasn't long before the internet saw the first Social Networking Sites (SNS). The most notable SNS that were founded in the early 2000s include: MySpace in 2003, Facebook in 2004, Bebo in 2005, and Twitter in 2006.

As the internet became more accessible, online communities began to define social norms rather than daily physical interaction (Hallam, & Zanella, 2016), contributing to a massive societal shift. This shift was influenced by "unprecedented access to information because sharing and spreading information [had] never been easier" (Lanzing, 2016, p.85). Commonly, research surrounding the use of user data is directly related to user response to ad personalisation (O'Donnell, & Cramer, 2005; Park, Campbell, & Kwak, 2012) and law policies (Mayer, & Mitchell, 2012; Roderick, 2014). The wide use of targeted advertising enables the personalisation of user experience online and ad-funding. This targeted advertising, in turn, facilitates free user access to online content (O'Donnell, & Cramer, 2005; Mayer, & Mitchell, 2012; Larose, & Rifon, 2007). While many users appear comfortable with

ad-funded content (O'Donnell & Cramer, 2005), many are not comfortable with the quantity or accuracy of the personal information gathered outside of their control (Roderick, 2014).

Over the past decade, the privacy paradox has only gained attention as it remains a cultural issue online. However, the focus of this field remains on either SNS or e-commerce. This places particular attention on the effects of the privacy paradox rather than the causes. However, the PP is not limited to only SNS and e-commerce sites online, but rather exists as a broad issue that affects all services online. Park et al. (2012) stated that users experience privacy concerns across most online services offered. To protect privacy across the range of online services, users must both understand how to protect their PII and how to maintain this behaviour across all services.

As current research has focused on investigating user opinions (Hallam, & Zanella, 2016; Liao, Liu, & Chen, 2011; O'Donnell, & Cramer, 2005; Park et al., 2012), little has been proposed to remedy the PP directly. User motivation is a significant deciding factor as to whether or not the PP remains without solution (Hallam, & Zanella, 2016). This is due to two main behaviours: internet users often neglect to maintain privacy protection behaviours across all online services; and, users can neglect to dedicate time to read and understand website Terms of Service, aka Terms of Use, or Terms and Conditions (Hillman, 2005).

It is evident that a vast amount of internet users experience concerns for PII protection online. As a result, majority of research focuses on SNS and e-commerce services (Hallam, & Zanella, 2016; Larose, & Rifon, 2007; Park et al., 2012) as they are the most prevalent service types that require user PII disclosure in order to use.

Understanding user privacy concerns (and how users act on these) has been a main aim of previous research. Current literature agrees that while users indicate they have privacy concerns for PII, many do not act on these—even when they understand how (Hallam, & Zanella, 2016; Larose, & Rifon, 2007; Liao et al., 2011; O'Donnell, & Cramer, 2005; Park et al., 2012). This is partially due to the nature of some online services, such as SNS, where users privacy protection is complicated by other users disclosing the PII of those they know without consent. This behaviour makes “the management of privacy, both for the individual and for organizations, extremely complex” (Dwyer et al., 2007, p. 3). While some literature argues that the privacy paradox is not a paradox at all and is instead differences in perspectives on privacy (Utz & Krämer, 2009), extensive research solidifies the existence of the privacy paradox as a social issue that has developed through the years since the beginning of the internet.

While online services that request or require the use of PII disclosure provide privacy policies that outline the treatment of this data (CalOPPA, 2016), it is the internet user's responsibility to read these legal policies. This area of self-motivation allows for the PP to flourish as the perceived benefits do not outweigh the perceived costs of investment (Hallam, & Zanella, 2016; Hillman, 2005). Factors that contribute to the perceived cost of reading these legal documents include (but are not limited to) document length and complex language (Larose & Rifon, 2007); factors that increase the amount of time required to read and understand the content.

The revelation that many internet users do not read website ToS and PP highlights that there is a potential correlation between their online behaviours of PII disclosure and lack of privacy protection. This idea links together two fields of research—PP research and ToS comprehension and retention research—through the identification of many overlapping factors. These factors include the lack of knowledge many internet users hold about the treatment of PII online, and how to better protect this information.

Rogers (1975) protection motivation theory states that if a perceived threat (such as getting hit by a car while crossing the road or a breach of privacy while browsing on the internet) creates a fear response, responses include motivation to prevent this threat from occurring. Youn (2009, p. 2) states that “self-efficacy, that is, individuals’ belief in their capacity to successfully carry out an action, is crucial in explaining protective motivation” in support of Rogers’ (1975) theory. These studies highlighted the importance of user awareness of their own vulnerability. If the user understands where a potential threat could exist, they are more likely to take actions in order to protect themselves from this perceived threat.

To improve privacy protection online, users benefit from a general understanding of how the internet works. However, it is more important for users to understand what happens to the information they self-disclose online. Internet users are widely aware that “personal data [is] oftentimes gathered in exchange for rewards” (Park et al., 2012, p. 1), however the extent of data collection is commonly not known. Most internet users understand and accept that personal data is collected and used for direct behavioural marketing (commonly called “targeted advertising”), however the amount collected and other uses largely remain unknown to users (O'Donnell, & Cramer, 2005; Wu, Huang, Yen, & Popova, 2012).

Website ToS have an important role ensuring users are informed on the practices of the service (Wu et al., 2012). ToS and PP contain all the information websites are required to disclose about how they use/repurpose user information, what data they collect, and who has access to what information (PII or otherwise). Country specific rules can affect ToS and PP regulations. To cater to different cultures, these documents are adjusted

to the privacy concerns of the population in each region when practicable (Bellman, Johnson, Kobrin, & Lohse, 2004; Wu et al., 2012). Although some complications occur through regulation conflicts when information is transferred from one country to the next, systems such as Fair Information Principles are in place to ensure user information is treated in adherence to the regulations specified (Bellman et al., 2004).

However, with most internet users overlooking ToS and PP documents (Hillman, 2005), it appears only natural the privacy paradox phenomenon would flourish despite the efforts made with current regulation practices. Although users often state they are concerned for their privacy online because they do not know what happens to the information they self-disclose, they often dismiss website legal documents (Hillman, 2005). The act of dismissing ToS and PP only allows the privacy paradox to continue existing, as this is the method online services employ to inform users of the treatment of user data.

If users do not read ToS and PP documents, they have an increased lack of awareness of the use of the PII they self-disclose (or disclose about others) on the internet services they engage with. Rogers (1995), and later, Youn (2009) both support the theory that user motivation is essential for inspiring protection behaviours. These studies also highlight the importance of users needing to be aware of their own vulnerability, stating that if they are aware of potential threats, they are more likely to take action to protect themselves. This emphasises the importance of internet users informing themselves of policy contents in order to understand how to protect PII on the services they engage with online.

With most studies in these research fields focusing on the effects of privacy paradox on SNS and e-commerce, or user willingness to read online legal documents respectively, there remains a gap in connecting these fields. Both research fields suggest providing more accessible ToS and PP will allow users to inform themselves of the treatment of their PII online. This would be beneficial to remedying the issues currently faced by both fields, as it could inspire privacy protection behaviours. Focusing on the effects of the problem rather than the causes has lead to vast knowledge on the social impact of this issue, and the correlating impacts on online service acceptance in the communities (Larose, & Rifon, 2007; Liao et al., 2011; Hallam, & Zanella, 2016; Mayer, & Mitchell, 2012; O'Donnell, & Cramer, 2005; Park et al., 2012; Roderick, 2014). However little research has been conducted to explore practical solutions to this issue.

Targeting gaps in user knowledge should hold more focus in literature as a means to eradicate the privacy paradox, with a redesign of ToS and PP documents being the initial step to take. These documents serve as the main means for digital contracting and awareness for internet users, but users are

too ready to dismiss the significance of reading and understanding these. This research will place focus on website legal documents as an entry point to propose a potential solution to a cause of the privacy paradox, and investigate the possibility of an informal redesign of these documents to attract online user attention to the content.

Due to the digital nature of the privacy paradox phenomenon, this research also considers the delivery and development methods for the potential solution it aims to propose. Due to the ever changing nature of the internet, it was important for the technology investigations to reflect this adaptability. While standard code libraries were considered, research shows Machine Learning (ML) technology is more efficient and adaptable when manipulating large datasets. ML is a subfield of computer science research which separated from the artificial intelligence subfield to focus primarily on solving practical problems.

Arthur Samuel (1959), suggests the use of ML for tasks that would otherwise require elaborate programming effort as this technology allows computers to learn and problem solve in a time efficient manner. ML systems can learn from data, improving with experience and allowing for a refined prediction model to develop, based on its learning (Bell, 2015; Jordan & Mitchell, 2015). This technology is primarily used when a process uses large datasets that need to be sorted algorithmically (Alpaydin & Bach, 2014). Automation systems often implement ML trained models as they allow for the development of adaptable, sustainable systems that make output predictions based off large sets of data without constant human interactions. ML models require human input during the dataset preprocessing, and the training phases (this includes the refinement of the model).

Combining the heavily theoretical and behavioural research of the privacy paradox and ToS/PP fields with adaptive code models to propose a design solution is essential. The complexity of the societal issues covered in this research are largely dictated by the current societal climate, and it is important that the research reflects this throughout the investigations inspired by these investigations into previous works.

PRECEDENT ANALYSIS

Investigating projects that aimed to address user interaction with ToS/PP and the readability of these documents were evaluated to understand the successes and failures of previous designs. This allowed the proposed solution published in this thesis to incorporate the successful elements of these designs, while building past the failings or flaws these designs experienced. This also allowed investigations into potential competitors for the proposed design tool, and helped prevent unintentional plagiarism of ideas currently available.

Figure 1: Terms of Service; Didn't Read Website home page

Figure 2: Terms of Service; Didn't Read Website Ranking for Google

TERMS OF SERVICE; DIDN'T READ

Originally Hugo Roy started the development of the 'Terms of Service; Didn't Read' (ToS;DR) project in June 2012, however since then it has been developed by a growing number of people. This project is designed to manually summarise website ToS for easy consumption by internet users. The concept was proposed during the 2011 Chaos Communication Camp near Berlin, with the overarching idea to make website ToS (and other "cloudy" services) user-readable ("Terms of Service; Didn't Read," n.d.). It was inspired by the "Privacy Icons" (Aza Raskin) and "TOSBack" (Electronic Frontier Foundation) projects.

This project offers a multifaceted browser extension that not only rates websites ToS by a classification system (ranging from class A as fair terms, to class E as seriously concerning) but also summarises the ToS so the user does not have to read the full legal document ("Terms of Service; Didn't Read," n.d.). These summaries are split into categories and provide a small summary on each topic of interest (with each of these topic getting their own rating – good, bad, blocker, or neutral). These summary categories are:

- Waiving user right
- Business transfers
- Changes
- Notice of changing terms
- Cookies
- Scope of copyright license
- Data portability
- Governance
- Guarantee
- Jurisdiction and governing laws
- Law enforcement and due process
- Law and government requests
- Right to leave the service
- Logs
- Ownership
- Personal data
- Easy to read
- Suspension and censorship
- Third parties
- Anonymity and tracking
- User choice
- User information

However, as these summaries are created manually, they lack the ability to adapt as ToS are updated (instead, they require manual inspection and re-summarising to ensure the summary still correlates to the updated version). Due to the large amount of summary categories, the resulting summary does not provide much solution minimising the length of these documents. Rather, this summary succeeds in providing content that is more readable.

Figure 3: Terms of Service; Didn't Read Collaboration with Duckduckgo tool (tested on the website www.google.com)

TERMS OF SERVICE; DIDN'T READ 2018 RELAUNCH: COLLABORATION WITH DUCKDUCKGO

After a recent collaboration with the anti-tracking browser service provider company DuckDuckGo, the ToS;DR project was relaunched, allowing for a revival of the project. This has helped provide the ToS;DR project a direct feed into the mobile browser version of DuckDuckGo as a feature, providing live feedback to the users seamlessly integrated into the browser service (“Privacy, simplified. — DuckDuckGo Browser Extension & Mobile App,” n.d.). This service also offers an updated version of the original browser extension for laptop and desktop based users who set DuckDuckGo as their default search engine. This collaboration employs the ToS;DR API that was first released in 2014. The DuckDuckGo add-on checks for updates every 24 hours to ensure the data displayed to users remains up-to-date with the ToS;DR versions and remains as updated as possible to provide accurate summaries for internet users.

Figure 4: Polaris website: showing a summary of *www.google.com*

Figure 5: PriBot: Showing a conversation based off the Privacy Policy from the website *www.google.com*

POLISIS

Released October 2017, Polisis was developed by a team of six researchers and developers over the course of twenty months and features two variant outputs.

This project was developed in response to internet users dismissing or becoming frustrated with the process of reading website legal documents, especially with the popularity of the Internet of Things (IoT) and small screen smart devices. To encourage user attention to PP online, the Polisis team developed a dual purpose output. The first is an automated framework for policies analysis, Polisis (see Figure 4). Polisis uses machine learning to automate the process of breaking up website PP into topic specific categories, allowing the user to see what the company collects, what is shared, and more (“AI-powered Privacy Policies,” n.d.). It eliminates the requirement to read a full PP and instead allows the user to understand what they agree to online through visualisations and extracted sentences. This breaks the policy into easily digestible paragraphs pre-organised into the topic themes for easy consumption for the user.

The second is a conversational privacy bot called PriBot (see Figure 5) which is designed to deliver the contents of the PP through a two-way dialogue between an AI computer agent and the internet user. While this design was first developed as a by-product of the 2016 chatbot “hype”, it also enabled the design cross platform compatibility and was motivated by its ability to translate towards conversation-first devices, such as Amazon Alexa and Google Assistant.

PriBot answers any questions an internet user has about the websites PP in real-time (“AI-powered Privacy Policies,” n.d.). While this was originally actually the starting point of the project, the focus shifted to the Polisis interface after two things became apparent. Firstly, PriBot is only effective if the user interacting with it has specific questions to ask about the PP (meaning this still requires a lot of direction from the user). And secondly, that during the development process of building this design, the team developed a general system that can automatically analyse PP documents through the use of machine learning. This final realisation was the spark that ignited development of Polisis.

Figure 6: Comparison between standard Facebook website home page and this page with the Gate Peepin tool active

GATE PEEPIN

Gate Peepin' is a Firefox plugin developed by Linda Hilfling that was launched on the 18th of December 2008 as part of the 'Speaking Out Loud' exhibition ("Linda Hilfling Gate peepin | NIMk," 2008). This project was developed as an artistic intervention allowing internet users easy access to ToS, by forcing content of ToS directly into the website. Rather than requiring internet users to find and read website ToS, this plugin finds common words from the website ToS and the front end of the website the user engages with. After finding these words, it then injects the ToS content into the front end so the user can "peep" into the otherwise long, hidden legal documents they agree to online (Gatepeepin, n.d.). This plugin was developed as a tool to alter the way users browse the internet, giving them a different online experience (see Figure 6).

Beyond the basic toggling on and off of the plugin, this project also offers users with a multitude of customisation features to further personalise their online viewing experience. This customisation includes aesthetic adjusting of the injected text, and further customisation aspects allowing for changes to the text filtering to enable the plugin to look for specific content to inject from the ToS.

Hidden within the plugin, this tool also provides a small summary of ToS including average time predicted for a user to finish reading the document, and a list of key words/phrases and the number of times they occur throughout the ToS. However as this feature is not the main focus of the tool, it requires more user engagement and interaction to access this section of the plugin. This results in this feature being easily overlooked or users not being aware it is a feature of the plugin at all.

Figure 7: Clickwrapped website homepage

Figure 8: Clickwrapped summary on the website "Google"

CLICKWRAPPED

This project, first created in August 2012, was written and maintained by Andrew Nicol until 2016 when the project was last updated (“Clickwrapped,” n.d.). Although intended to grow in size, this project remained small in scale; only addressing fifteen websites.

This project takes the website ToS and gives them a score out of 100 (see Figure 7). This score is used to indicate how well the website respects user rights (a higher grade translate to mean the website treats their users with more respect). These scores are calculated through a ranking system between four categories, with certain questions being asked per each category (“Clickwrapped,” n.d.):

Data use:

- What data does the site collect?
- What can the site do with content you post?
- Does the site get more rights to your content than it needs?

Data disclosure:

- When can the site disclose your information to others?
- Does it tell you if the government wants your data?
- Is it transparent about government requests?

Amendments and Termination:

- Does the site give you notice when it changes its terms?
- Under what circumstances can your account be terminated?
- Does the site let you take your data to another service?

Miscellaneous:

- What else should you know about the site’s legal agreements?

While this project provides a concise summary for each of these categories on the website for users to read (meaning they are not required to read the full ToS and PP themselves), it is not clear which version of these legal documents has been summarised (see Figure 8). This results in some confusion as to whether these summarises are out-of-date or how long it has been since they were last updated (however, the website has not been updated since 2016 so the accuracy of these documents cannot be confirmed). There is also some ambiguity in this project as to the exact calculations of the scores given to the sites. While the Clickwrapped project offers some explanation as to why the scores have been given, these are not numerical justifications resulting in the number scores feeling disconnected and less valid.

Figure 9: Privacy Guide application showing a comparison before and after loading a website URL (www.unileverprivacypolicy.com)

PRIVACYGUIDE

Published in April 2018, PrivacyGuide is a project developed by a team of five as a direct response to the emergent EU General Data Protection Regulation (GDPR). The overarching goal of this project was to “take the requirements stated in the GDPR as a basis and provide a privacy policy benchmarking tool that considers privacy aspects and through the use of risk indicators [PrivacyGuide] draw[s] users’ attention to the most relevant parts of the privacy policy” (Hofmann, Nakamure, Kiyomoto, Serna, & Tesfay, 2018). Using the GDPR guidelines, this project uses a “conceptual grouping of privacy aspects” (Hofmann et al., 2018) to create a supervised machine learning model to “classify the lengthy privacy policies and employ a three scale risk based categorisation” (Hofmann et al., 2018) to simplify PP interpretation for internet users.

This is a stand-alone tool that functions as users input the URL of the PP they want summarised (see Figure 9). The application summarises the defined PP and shows the results in a visually clear scale of green-yellow-red (green meaning best and red meaning worst). The project, still in the earlier stages of development, has yet to undergo usability or user adoption testing so it remains unclear if the transfer from the browser to a separate application is an effective method of delivery for this tool to users.

Overall, PrivacyGuide is very effective in proving that machine learning techniques can be used in this field of research. While the application is not yet readily available for users, the demo is a promising proof of concept.

RESEARCHER'S NOTE:

PrivacyGuide was published during the course of this research, and proved to be an immediate competitor as the overall design of the system is very similar to that which this research also undergoes. The development and publication of this project during the course of this thesis also solidifies the relevance of this field of research in the current social climate.

METHODOLOGY

This research uses a mix of traditional and contemporary research methods throughout the contextual research and development of a final design output. These methodologies provide ways to investigate and develop research that is situated within the research field and builds on the previously conducted research. The combination of traditional and contemporary methods allows the research to explore design through reputable and established tradition methods, while also using contemporary methods specifically designed to address contemporary issues such as those this research is investigating.

CONTENT ANALYSIS (INDUCTIVE)

Content analysis is an important method for communication research as it encourages analysing data within a specific context, with particular attention to the attributes of a particular group or culture (Krippendorff, 2003). This method allows for objective exploration into all aspects of the investigated text/content and establishes the inquiry focus.

This research employs the use of content analysis methodology to analyse website ToS and PP. This is used to identify important themes and common points, locating not only the important themes of each websites individual policy documents, but also the significant themes that are consistent across the multitude investigated during the research process. The conducted analysis is inductive, allowing “categories or codes [to be] derived from a systematic reading of a sample set of the materials to be analyzed, gradually establishing the categories that will be used for subsequent analysis of all the materials” (Hanington & Martin, 2012, p40).

This is important for understanding the repeating elements of these policies and identifying the recurring themes and patterns. These elements provide insight into the main themes of the legal documents prior to design output prototyping.

CASE STUDIES

The use of case studies enables the research to provide in depth details to explain real world examples, without delving into a scope this thesis cannot facilitate. It enables explanation through real world examples, showing proof of the situations that the research context is founded on. Using case studies provides a smaller, more specific example appropriate to the scope of the research while also providing key contextual insights.

“The case study method focuses on gaining detailed, intensive knowledge about a single instance or a set of related instances” (Hanington & Martin, 2012, p28), allowing for an in-depth analysis of cases within the context of this thesis. This enables exploration of existing phenomenon for the purpose of better understanding the research field (Hanington & Martin, 2012). This method allows for the drawing of conclusions and making comparisons of pre-existing situations and designs for a more informed and considered research output.

This research employs case studies to allow for exploration into how ToS and PP documents and the regulations of these vary per country. This allows for explorations into the different considerations that factor into the forming of these documents by the service provider and encourages discussion on how these regulations are reflected in the documents themselves.

QUESTIONNAIRE

As a survey tool, a questionnaire allows for asking a number of people to answer a list of questions within the same timeframe. This tool is used for larger scale collection of generalisable datasets, often used to establish a situation or context and the identification of public opinion. Questionnaires can have questions that collect qualitative data, quantitative data, or a mix of both.

The use of an anonymous online questionnaire will be employed for the purpose of this research in order to check any assumptions introduced to this research. This involves checking that issues addressed in older research is still ongoing today (and has not been resolved or has stopped existing with the introduction of new technology), and any personal assumptions I, as the researcher, has brought to the research (Hanington & Martin, 2012). This method allows for checking these assumptions and providing a valid ground for the research to build upon, helping minimise the bias(es) brought to the research.

Conducting a questionnaire alongside this research also enables me to place this research within a smaller, local context while also situating it within the broader context of the research field. This is both important for establishing the overall context and applicability of the thesis on a larger scale as well as a situating it within the same context the research exists immediately within. This also ensures that there is an easier identification of the assumptions I

brought to the research, which require validation or dismissal during the process of this research.

CONFIGURATION

Configuration as a research method encourages dismantling the problem/issue the research is investigating to value each influencing factor or parties equally in the creation of a solution (Suchman, 2013). This encourages investigations into the socio-technical nature of the research while also focusing on the sustainability of the developed output.

Valuing each influencing factors of the potential output and research issue encourages the research to successfully identify the motives of the parties involved and what they will value most in the final product (Suchman, 2013). This prediction allows for the research to shape the design output estimating the expectations of the parties, so it can better cater to these. Being conscious of the involved parties as well as the technical requirements of the research directs the design output allowing the research to balance these equally, to best serve all factors without catering to one party more than others. This creates an output that better caters to everyone while still providing a solution to the issue proposed by the research.

Given the multifaceted nature of the issues this thesis aims to address, it is important to identify all the parties that will be impacted by the proposed design output of this thesis. This will enable for a better informed design output that does not impose a bias either in favour of the service provider, or against. This will, in turn, aim to provide the user of the design output a neutral indication that they can then make a more informed decision on.

PROTOTYPING

The design development process will follow an iterative prototyping approach, allowing for rapid experimentation of the final design output. This allows for the output to trial different technologies to establish which is the most effective method for completing the intended output functionality (Hanington & Martin, 2012). This also allows for quick addition of potential features users expressed they'd prefer during the questionnaire circulation process that will be completed prior to this step in the research.

This also allows for quick addition or removed of features. And provides an insight into the multiple ways the output could be built, with the final output focusing on the most efficient of these.

REGULATION AND LEGISLATION

During the literature review, the user perspectives of the privacy paradox phenomenon were covered in detail and considerations into regulations were briefly mentioned. In this section, the perspectives of online services providers were explored. This largely involves investigating the regulations and legislations that these companies must comply with. These investigations build on the work of Bellman et al. (2004) and Wu et al. (2012), which explored country-specific governance and general regulations of online legal documents.

It was important for this research to cover the regulations that govern online legal documents. This enables informed observations on the user treatment of these documents, and suggestions for changes that services could implement while still complying with the regulations governing these policies.

Understanding the legal issues that impact the ways these policies are treated in general and on a per country basis provided the legal context that this research is situated within. Investigating the regulations online service providers must comply with was also important for understanding their perspectives on the privacy paradox. This encouraged considerations into how these regulations may impact the potential design output.

CASE STUDIES

Regulations and legislations vary across the globe due to multiple factors such as local culture and political climate. Evidence shows that while online regulations reflect their local climate (Utz and Krämer, 2009), online services which cater to residents in different countries are also subject to the regulations of the users country. This introduces a level of complexity to ToS and PP document contents as they must adjust to comply with rules which reflect cultures outside of their immediate environment.

NEW ZEALAND

New Zealand internet regulations primarily focus on the safety of users online. These regulations are in place to monitor for harmful or illegal activities – with particular attention to distribution or trace of objectionable or restricted material (“Censorship and the Internet - dia.govt.nz,” n.d.). Other behaviours that are monitored include those which are considered detrimental to the safety of the nation; these include: gambling, spam, money laundering, personal security, and any objectionable or restricted material (such as child pornography).

Department of Internal Affairs

Primarily, the management and enforcement of the laws currently in effect in New Zealand fall under the power of the Department of Internal Affairs (“Censorship Online Safety - dia.govt.nz,” n.d.). These laws include (but are not limited to):

- Films, Videos, and Publications Classification Act 1993 (updated in 2015)
- Gambling Act 2003
- Unsolicited Electronic Messages Act 2007
- Anti-money Laundering & Countering Financing of Terrorism 2009
- Private Security Personnel and Private Investigators Act 2010

While these laws include many acts that are made easier by the existence of online services, predominantly services offered to New Zealand citizens based internationally, the reach of the Department of Internal Affairs is limited to enforcing these laws within New Zealand based services (“Contract and Commercial Law Act 2017 No 5 (as at 14 July 2017), Public Act – New Zealand Legislation,” n.d.).

UNITED STATES OF AMERICA (USA)

The United States of America has one of the largest contributors to online consumer and commerce presence across the globe. As a result, USA has a large variety of online regulations that affect not only the residents of the country, but also influence international data regulation.

Federal Trade Commission

The Federal Trade Commission (FTC) was established in 1914 with the implementation of the Federal Trade Commission Act law (Federal Trade Commission, 2013). The FTC is an American agency with the mission to protect consumers and promote competition. The FTC regulates the data protection of all American consumers and works with international agencies to further protect these consumer in the international marketplace. This regulation includes more than 70 federal and state laws.

The FTC conducts checks to ensure companies or people do not violate the law, and educate these companies and people on their rights and responsibilities in the consumer marketplace. Monitoring company practices to ensure the market caters to consumer preferences, and not unfair or illegal practices also falls under FTC jurisdiction.

NOTABLE FEDERAL AND STATE LAWS CURRENTLY IN USA INCLUDE:

1. The Children's Online Privacy Protection Act (COPPA)

COPPA is an internet law that was passed by congress in 1998 to protect the privacy of children under 13 years of age. "Young children" under 13 years of age were decided to be of higher risk of PII abuse online. Any website that functions in agreement with the laws defined by the FTC that collects information from children within the defined age bracket (under 13s) are required to comply with COPPA.

To comply with these specific rules, "a Web site operator must include in a privacy policy, when and how to seek verifiable consent from a parent and what responsibilities an operator has to protect children's privacy and safety online" (COPPA, n.d., para. 1). The FTC determines if a website targets children by evaluation factors such as; visual or audio content, subject matter, subject model age, language use, whether advertising targets children, child-oriented features (such as animated characters), and whether the site offers information about the intended audience. The role of the website in consideration to the collection and maintenance of the information is also evaluated by the FTC to determine if the service entity is an "operator".

RESEARCHER'S NOTE:

The implications for younger (under 18) audiences did not factor into the development of this thesis as it was out of scope to include considerations specific to minors. As a result, the design output is designed for users 18 years and older.

2. The California Online Privacy Protection Act (CalOPPA)

First initiated in 2004, this is a state law (not a federal law) that requires any commercial websites that collect PII from California's residents to provide a conspicuous link to the PP document on their website. The scope of this law is very broad, as it encompasses companies and web servers that are not located in California specifically, but rather applies to all websites that are accessible by California's residents ("CalOPPA," 2016).

This law states that a website must post information regarding the method of PII data collection and use. This includes details about third-party entities with whom the service provider shares this information with. The service provider must also provide a method for the user to review and change any of the collected PII through the website or online service ("CalOPPA," 2016). Disclosure of other parties collection permissions on the service must also

be present in this policy, including the definition of whether this is across third-party sites, over time, or whether the original service provider engages with this collection.

EUROPEAN UNION

The European Union (EU) includes over 20 member states. Across all member states there is a standardised system of regulation and laws, developed as a single internatised marketplace. This standardisation allows for free exchange of people, services, and capital.

This standardised system includes many regulations that aim to become the forefront of internet user empowerment and ensuring privacy protection online.

The EU General Data Protection Regulation (GDPR)

Approved by EU parliament on the 14th April 2016, the GDPR replaced the Data Protection Directive 95/46/EC on the date 25th May 2018. The regulation “seeks to harmonise the protection of fundamental rights and freedoms of natural persons in respect of processing activities and to ensure the free flow of personal data” (European Commission, 2016, p. 1), empowering EU citizens data privacy. GDPR was developed to establish one set of data protection rules across Europe. However, many organisations outside of the EU are also subject to complying with this regulation, as any service that collects data about a EU citizen must also comply.

Penalties for companies found to be in noncompliance will face the greater of either ten million euros or four percent of the organisation’s annual revenue for violations. Violations include:

- Record-keeping
- Security
- Breach notification
- Privacy impact assessment obligations

These penalties are doubled if violations are related to legal justification for processing, lack of consent, data subject rights, or cross-border data transfers.

Companies are required to “implement appropriate technical and organisational measures to ensure and to be able to demonstrate that processing is performed in accordance with this Regulation” (European Commission, 2016, p. 47). This requires data protection elements to be implemented into services and products from the early stages of development “by design and by default” (European Commission, 2016, p. 15). These elements act as safeguards and are designed to reflect the type of data collected.

One of the most significant parts of the GDPR is the requirement for consent from the individual whose data is collected. This regulation defines consent as “any freely given, specific, informed and unambiguous indication

of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her" (European Commission, 2016, p. 34). As part of this consent, companies must be able to prove how and when consent from each user was acquired, and data must only be collected for an explicit purpose. As part of this consent, users can also withdraw their consent at any point and, unless the data is still required for the initial reasons for collection, it must be deleted.

The service provider is only allowed to collect information for legitimate purposes, and its use must be made clear to the user. To increase transparency, companies must give users full access to their data, and information about how this data is processed. This information must be clear and easily understandable. Following this trend of increased transparency, companies must also report any security breaches "leading to the accidental or unlawful destruction, loss, alteration, unauthorised disclosure of, or access to, personal data transmitted, stored or otherwise processed" (European Commission, 2016, p. 34).

EU Directive on privacy and electronic communications "EU Cookie Law"

This legislation requires all websites to gain consent from users to store or retrieve information on any device (computer, tablet, or smartphone, etc). Initially implemented as a means for privacy protection, it also acts as a means to increase user awareness of information retrieval methods and use, and allows the users a choice of allowing this. This legislation increases user rights by allowing them to deny the use of cookies (and similar technology) that would otherwise reduce their online privacy ("The Cookie Law Explained," n.d.).

The extent of this laws requires all websites inside the EU to comply, as well as any websites that target EU citizens.

SOCIO-TECHNICAL ASPECTS

After establishing the different perspectives of both the users and the service providers separately, it is important to consider what can occur when these two sides interact. At their core, the two perspectives are at odds as they treat their own perspective as more important than the other. This is especially important when PII is disclosed on a service.

Users want the ease of many free online services, and service providers want to be legally protected and still run a sustainable business model. These two viewpoints differ when users want full privacy protection and online service providers sustain their services through business models founded on advertising. Internet users like to assume that their information is kept safe after they disclose it online. However, due to the nature of advertising

focused business models, the repurposing of user information is often how “free” online services are sustained (O’Donnell, & Cramer, 2005; Mayer, & Mitchell, 2012; Larose, & Rifon, 2007).

It is important for the treatment of PII online to be transparent for the benefit of both parties—user and service provider—as it allows for stronger trust to form. However, it is also important of the service provider to structure their ToS and PP documents so that they are unambiguous and clearly state their intentions. This can result in a complicated document, that is not necessarily beneficial to the user. The collection of user data to facilitate targeting advertising is one example of an interaction between user and service provider that is allowed through strict regulation. However, the interaction between user and service provider is not always a smooth, compliant interaction. Below are two example situations which have the ability to influence greater transparency in the method service providers gain consent from users and use their PII.

FACEBOOK AND CAMBRIDGE ANALYTICA DATA “SCANDAL”

The Facebook and Cambridge Analytica data “scandal” came to the attention of news media in March 2018; it highlights the importance of transparency between service provider and user in regards to data use. This data “scandal” was an extreme example of data mishandling by the service provider ‘Facebook’ that was originally allowed through the ToS which users agreed to (it has recently been amended so this could not occur again). This example shows how easily PII can be abused online.

Without proper attention to the permissions users give services and applications (as outlined in their ToS users agree to), users can open themselves up to the chance of data abuse such as this.

Timeline

2010

- Facebook launched their ‘Open Graph’ API to third party application developers.
- Allowed external developers access to Facebook users personal data and the personal data of their Facebook friends. This data included (but was not limited to): name, gender, location, birthday, education, political preferences, relationship status, religious views, and online chat status.

2013

- Cambridge academic Aleksandr Kogan and his company Global Science Research published the application ‘thisisyourdigitallife’ on Facebook.
- This app created a psychological profile of the user.

- The application allowed the harvesting of data from the (approximate) 300,000 users who took the test and their Facebook friends data.

2014

- Facebook changed their API rules to limit developer access to user PII.
- Third parties could no longer access Facebook friend information indirectly.
- These changes were not retroactively imposed – allowing Kogan to keep the data he improperly acquired.

DEC 11, 2015

- Cambridge Analytica allegedly helped Ted Cruz with his presidential campaign reported The Guardian (Davies, 2015).
- The report also suggested that this campaign used psychological data based off Facebook data to gain an advantage over political rivals (Davies, 2015).
- Facebook claimed this was the when they became aware of a “data leak” which prompted them to demand Cambridge Analytica and Kogan delete the user information.

2016

- Donald Trump’s presidential campaign invested heavily in Facebook advertisements (with the help of Cambridge Analytica).
- Mark Turnbull (a Cambridge Analytica senior executive) later told British Channel 4 in an undercover interview that the firm was responsible for the ‘Defeat Crooked Hillary’ Facebook video campaign (“Exposed,” 2018).

MARCH 17, 2018

- The Guardian and The New York Times reported 50 million Facebook profiles were harvested by Cambridge Analytica (this was later revised to 87 million).
- Whistleblower Christopher Wylie revealed the alleged practices to these news firms.
- Wylie claimed data sold to Cambridge Analytica was used to develop psychographic profiles to deliver pro-Trump material to potential voters (Cambridge Analytica denied that this information was provided by Kogan’s app).

MARCH 20, 2018

- Federal Trade Commission (FTC) opened an investigation into Facebook to check if they violated a prior settlement reached between the U.S. government agency and Facebook in 2011 over user privacy protections.

MARCH 21, 2018

- Zuckerberg responded to Cambridge Analytica news, announcing third party developers can no longer access user data after three months of inactivity, and have reduced the amount of information third parties can access from users.

APRIL 10, 2018

- Zuckerberg appeared before the U.S. Senate Judiciary and Commerce Committee to answer questions surrounding the validity of the claims Wylie made surrounding Cambridge Analytica's use of Facebook's user data ("Facebook, Social Media Privacy, and the Use and Abuse of Data | United States Senate Committee on the Judiciary," 2018).

APRIL 11, 2018

- Zuckerberg appeared before the U.S. House Energy and Commerce Committee to answer questions aiming "to shed light on critical consumer data privacy issues and help all Americans better understand what happens to their personal information online" ("E&C Announces Facebook CEO Mark Zuckerberg Will Testify April 11th," 2018).

MAY 22, 2018

- Zuckerberg appeared before the European Parliament to answer questions surrounding topics of data policy, terrorist content, disinformation, monopoly power, and more ("EP Conference of Presidents with Mark ZUCKERBERG - founder and CEO of Facebook| European Parliament, 2018).

This event gained media attention in 2018, allowing for the crossover onto the GDPR adoption date. This increased the relevance of the GDPR in the eyes of the public as it became apparent how little many users understood about the permissions they give services they engage with.

GLOBAL GDPR ADOPTION

As the GDPR came into effect many online services struggled to comply with the new demands for privacy protection. The infrastructure and redesign of systems many companies previously employed was insufficient when compared to those the GDPR demanded.

On the day of regulation implementation (May 25th 2018), the newly founded European Center for Digital Rights organisation noyb.eu filed four "forced consent" complaints against Google (Android), Facebook, Whatsapp, and Instagram ("noyb.eu | My Privacy is none of your business," 2018). The complaints that were filed are as follows ("noyb.eu | My Privacy is none of your business," 2018):

- Google LLC (Android) filed with the French DPA (CNIL).

- Instagram (Facebook Ireland Ltd) filed with the Belgian DPA (CPP).
- WhatsApp Ireland Ltd filed with the Hamburg DPA (HmbBfDI).
- Facebook Ireland Ltd filed with the Austrian DPA (DSB).

If these complaints are upheld, the fines for each company could result up to 4% of their annual profit (maximum fine). The major issues these complaints bring to attention focus on user option to opt-out of consent and still interact with the provided service(s), choosing instead to protect their privacy. Another topic the GDPR has brought to attention, is “whether the processing of data for targeted advertising can be argued to be necessary for the fulfilment of a contract to provide services such as social networking or instant messaging” reported the Guardian (Hern, 2018, para. 10).

QUESTIONNAIRE

This questionnaire was conducted to determine if the results of previous formal and informal surveys are still applicable today. This questionnaire was also conducted to weigh the existence of privacy concerns within New Zealand internet users. The use of an anonymous online questionnaire allows for the checking of assumptions formed during the initial contextual research. These assumptions include:

- Internet users neglect to read Terms of Service and Privacy Policy documents;
- Internet users experience concerns for their privacy on the internet ;
- When experiencing concerns for their personal privacy online, internet users do not act on these concerns.

It is important for these assumptions to be checked, as they could prove to be pivotal for the course of the research as they will be used to shape the context of the design output.

While the assumptions checked in the questionnaire are formed after the contextual research, it is important to confirm that these are present in local society (in 2018). It also allowed for exploration into tangential issues and/or themes that factor into the research, and provided further insights into the topics this research aimed to cover. The conducted questionnaire focused on themes of:

- User understanding of treatment of information online.
- Privacy concerns experienced by internet users.
- How internet users protect their information online.

The main purpose, however, of this questionnaire was not to provide population representative results for future research to build off. Instead, this questionnaire was conducted to help situate previous research within the local geographic and help identify core themes to focus on through the design output development.

PREVIOUS SURVEYS

Popular opinions such as “nobody reads terms and conditions” and “I agree to the terms of service is the biggest lie on the internet” have risen in popularity as the internet grown more pervasive in society. To confirm or deny these claims, multiple surveys have been conducted in the past—by both news media and researchers alike.

Researcher Florencia Marotta-Wurgler conducted an investigation in 2009 that, as reported by Forbes, foretold that “just .11% of users click on a link to a site’s terms of service, about one in a thousand”, with the addition that “In the case where a company requires a customer to click a box that says ‘I agree’ to those terms and also offers a link, the number of users who click through is actually lower: just .07%” (Greenberg, 2010, para. 4). In the published research paper, it was concluded that “even under generous assumptions, it is hard to envision the probability that [end-user license agreements] are read, and understood, growing even to 1%” (Bakos, Marotta-Wurgler, & Trossen, 2009, p.35). This is due to the combination of users not clicking links to the policy documents and/or not spending enough time on these pages to realistically have read and understood them; with the average time spent on each policy page 59.4 seconds (Bakos et al., 2009).

This research opened the field for more investigations across the years to check/confirm if these findings persisted as consumers continued to engage with services online. Highlighting that issues surrounding users’ motivation to read ToS and PP documents were present previously also highlights the importance of checking if these issues have continued to the present day.

During an interview conducted by National Public Radio (NPR), Jonathan Obar—a researcher at York University who focuses on information and communication policy—was reported saying “It would take the average user 40 minutes a day to read all of – and that’s every day – to read all of the privacy and terms of service policies that we encounter related to the different services that we’re using all the time” (“Do You Read Terms Of Service Contracts?,” 2016, para. 16). During the interview, in-house social science correspondent Shankar Vedantam also proposed the question, “from the point of view of companies and policy-makers, the idea that these documents are eliciting informed consent is really a joke. They may provide legal cover, but how many people are really providing informed consent?” (“Do You Read Terms Of Service Contracts?,” 2016, para. 33).

This final design questions touches on some of the many changes the earlier mentioned GDPR (brought into effect two years after this interview) aims to challenge. The requirement of consent, and ensuring users are adequately informed are large components of the GDPR, with surveys such as this previous work proving that it is addressing issues which have persisted over the course of multiple years.

The paper Obar published in 2016 that was discussed in the NPR interview (“Do You Read Terms Of Service Contracts?,” 2016), co-written by Anne Oeldorf-Hirsch, employed the use of a fictitious social networking site. The investigation conducted requested participants to sign up for the fictitious site, with measurements relating to the attention participants gave ToS and PP. The overall results showed that “74% skipped PP, selecting ‘quick join’” and, for those that did read, the “average PP reading time was 73 seconds, and average ToS reading time was 51 seconds” (Obar & Oeldorf-Hirsch, 2016, p. 2). This research confirms that the earlier conclusions of Marotta-Wurgler’s research are still relevant in the online climate of 2016.

PARTICIPANTS

The questionnaire had a recorded 61 responses. Of these 61, 18 were removed as unfinished (“deserted”) responses and 20 were partial responses (these 20 included responses that the respondent had not completed some open qualitative questions asked). This left 23 fully completed responses.

The respondents were largely represented by females and the age range 19-24, with no respondents from the 35-44 or 55+ age ranges. As this questionnaire was conducted in 2018, this means the most common age group of respondents were born between the years 1994 and 1999. If you refer back to the earlier contextual research in this thesis, you will see that this is the first age group to grow up with significant internet integration into everyday society.

DISTRIBUTION

This questionnaire was distributed on the internet social networking platform Facebook. Participants were solicited through a Facebook “post” (Figure 10), with links to important information attached. These links included one link to the questionnaire (conducted through Qualtrics) and a link to the information sheet informing the respondents, in detail, to the purpose of this research questionnaire (see Appendix).

This questionnaire was distributed over two weeks in 2018 (first distributed on 28th February 2018, and closed for responses on the 9th of March 2018). The questionnaire was shared on my personal Facebook profile, inviting Facebook “friends” to complete the questionnaire at their leisure.

The distributed survey contained a combination of qualitative and quantitative questions. The survey consisted of a mix of likert scale quantitative questions, and open-ended qualitative questions. This provided the respondents with the opportunity to both answer some questions on a ranked scale and also gave the opportunity to explain some answers in further detail.

The Questionnaire was conducted to confirm or deny the claims of previous informal and formal surveys assessing the privacy concerns internet users experience while interacting on the internet.

Figure 10: The Facebook “post” used for distribution of the questionnaire

LIMITATIONS

Due to the scale of the questionnaire, this questionnaire was not intended to be interpreted as representative of the population (nor did it aim to be). The method of sharing on Facebook required all respondents to have a connection to the researcher. This included either directly as 'friends', or through 'friends' of Facebook friends who 'shared' the post created by the researcher (of which there were three associated friends who shared the post around their Facebook connections). This introduced an unintentional bias in the participant sample that completed the questionnaire (Babbie, 2013). However, as the questionnaire was kept anonymous and participants were not given rewards for completion, reactivity bias was minimised (Babbie, 2013). Complete anonymity also avoids potential hesitation or embarrassment respondents may feel when posed with questions they believed their response/opinion may be in the minority. Respondents want to protect themselves from ridicule, and can answer how they think they should answer rather than answer honestly if the questionnaire is not anonymised (Babbie, 2013).

The main bias imposed on this questionnaire, however, is more significant than the others. The questionnaire was only circulated on the social media platform 'Facebook'. This required all respondents to have a Facebook account in order to find the questionnaire. This bias introduces both a negative and a positive impact. The negative: it eliminates any participants which may be more protective of online privacy and thus refuse to create online social media accounts. The positive: it ensures all respondents have encountered both a websites ToS, and PP (the topic points the questionnaire focuses on).

RESULTS

There were three main themes that the questionnaire answers grouped within. The themes were:

- Privacy concerns
- ToS and PP knowledge
- Online behavior

Within each theme, there was a range of questions that assessed the multifaceted nature of human perception and behaviours. These questions have been summarised and organised into these themes below for identification of the traits within these themes.

ONLINE BEHAVIOURS

The majority of respondents engage in browsing the internet 26+ hours daily (see Figure 11). 26+ hours a week indicates a large range for variance outside of the range this questionnaire investigated, however it indicates that the majority (approximately 70%) of respondents engage with the internet on a regular basis (see Figure 12).

Figure 11: Questionnaire results; average hours spend browsing the internet

Figure 12: Questionnaire results; services actively engaged with online

Figure 13: Questionnaire results; types of information users are comfortable sharing online

Most respondents offered multiple answers when asked which services they engage with while online, with the average amount of services each respondent engaging with six to seven (see Figure 13). While there is no obvious “most popular” service, the top two included communication and entertainment. Excluding the “other” category (which required users to type a custom answer), the least common service still 26% respondents admitting to engaging with this service online.

This shows that internet users commonly disclose PII to a range of online services, often on a regular basis. This questionnaire suggests that the majority of internet users are exposed to multiple ToS and PP documents each time they engage with the internet.

When asked, respondents admitted that while engaging with a range of online services, they are open to disclosing a range of personal information. Most respondents answered that they are willing to share information from at least five of the defined categories, with multiple respondents stating they are comfortable sharing information from all categories (see Figure 13). The top information categories include those which can surface in everyday life (name and location), however the categories that were least likely for respondents to comfortably disclose online included more personal information (such as medical information).

PRIVACY CONCERNS

While most respondents answered that they experience at least a slight level of concern for their privacy online, no respondents claimed to experience no concerns (see Figure 14). 26% of respondents stated they were ‘very concerned’ or higher, meaning approximately a quarter of respondents experience high levels of concern for their privacy online. The main concerns stem from the fear that the respondent could experience financial loss or open themselves to spam/junk mail (see Figure 15). However, most respondents stated they have concerns that three or four different negative effects could occur due to their online behaviours, with multiple respondents concerned that all six above mentioned activities could occur to them.

The 4% of respondents that experience extreme concern for privacy online take the most elaborate measures for privacy protection online, including “regular password changes”, “2 step authentication”, and “a different password for every type of service”. To contrast, the respondents who stated they experience ‘moderate’ concerns for privacy online contained the majority of respondents who simply minimise online interactions and the limit the amount of information they disclose online, choosing to protect information by refusing to disclose PII online. Across all measures of concerns however, respondents stated that they employ “strong passwords” and ensure privacy settings are restricted. However, there was one outlier who stated they were ‘slightly’ concerned for their privacy, and thus “don’t take any actions” as they are “not too concerned”.

Figure 14: Questionnaire results; level of concern experienced by users online

Figure 15: Questionnaire results; types of events users are concerned about occurring while online

There appeared to be no link between the perceived personal ability to identify trustworthy sites online and the level of concern they felt while browsing the internet. The respondents which answered they experienced extreme concern for their privacy online all answered that they were ‘somewhat likely’ to identify trustworthy sites online. Respondent methodology for identifying these sites varied considerably, with one respondent relying solely on virus protection to tell them if a site was trustworthy, and another simply looking to check “if the address looks funny or you get a little warning don’t go there”. Also included in the ‘extremely concerned’ group was a respondent claiming to use eight different methods of identifying trustworthy sites (with some differences dependent on the type of service provided). These methods included “wide user base”, clarity “about the background or purpose of the site/service/creators”, and investigating the credibility of the name and/or affiliations.

In contrast to this, respondents who stated they experienced ‘slight’ concern for privacy online rely on “guts and intuition” (40%) or the appearance of a “padlock in the address bar” (40%), with only the remaining 20% considering other aspects of the site such as user base and site purpose.

TERMS OF SERVICE AND PRIVACY POLICY KNOWLEDGE

The questionnaire showed that more users read ToS than PP documents online (see Figures 16 and 17), however this could also be credited to many online services only having a ToS. With online services only providing a PP document when they collect or repurpose user PII, the frequency of reading these documents is not directly comparable. However, as the questionnaire moves into asking the importance of reading these documents, we can begin to see a disconnect between actual online behaviour and perceived good behaviour (see Figure 18).

When compared to the previous data, an immediate disconnect between behaviours and perceived good behaviours became apparent. Although none of the respondents previously stated that they always read ToS and PP documents, 17% still stated they believe it was extremely important to always read these documents before agreeing to them. Continuing with this trend, approximately 4% stated they believe it was not important to read these documents, a strong contrast to the 30% that never read these documents. This indicates that the actions internet users believe are correct online are not always the behaviours they take.

This research then compares this to how often they read the updated versions of these documents where applicable. The results show the majority of respondents do not read the updated versions, even though majority of respondents previously stated that they believe that the act of reading this document was important.

Figure 16: Questionnaire results; how often users stated they read website Terms of Service

Figure 17: Questionnaire results; how often users stated they read website Privacy Policy documents

Figure 18: Questionnaire results; how important respondents believe it is to read website Terms of Service and Privacy Policy documents

FINAL FINDINGS

Linking back to past research, the questionnaire conducted during the course of this thesis ultimately confirmed that the issues identified are still present. Internet users are still not providing informed consent to online services as they are neglecting to educate themselves properly to the information contained within the policy documents. All the discussed surveys and questionnaires highlight the leading cause of these issues is user failure to read online legal documents.

The important elements that were taken from the questionnaire included those surrounding internet user ToS and PP reading habits. This incorporated the justifications users gave as to why they do not read website legalese, the topics within ToS and PP that users stated they considered most important, and how important users felt it was to read these documents. These elements were then used to directly influence the developed design output, enabling the design to better cater to the issues it aims to address.

DESIGN IDEATION

PROCESS

The design process includes an iterative design approach, however this iterative process also followed an overarching step-by-step process. These steps are as follows:

- Concept Ideation
- Plan Functionality
- Build the Data Repository
- Plan UI/UX
- Create Functional Proof of Concept
- Train Machine Learning Model
- Combine Machine Learning Elements with Proof of Concept

Having a clearly defined process enabled adequate attention to be placed on each contribution towards the design output. This allowed for consideration into all aspects of the design to ensure the output reflected the goals of the research correctly.

ITERATIONS

The first step of the design creation process included ideation of the design, as lead by the research conducted through this thesis. This allowed the design output to grow and progress to reflect the research directly.

ITERATION ONE: READABILITY TESTING

Initially, this research was aiming to focus on the theoretical elements of this field rather than creating a practical design output. The first concept was aimed at investigating the reasoning behind why internet users don't read or understand ToS. These investigations focus on the readability of website ToS documents and how this may affect the target users ability to understand.

$$206.835 - 1.015 \left(\frac{\text{total words}}{\text{total sentences}} \right) - 84.6 \left(\frac{\text{total syllables}}{\text{total words}} \right)$$

Figure 19: *The Flesch Reading Ease formula*

Table 1: *The Flesch Reading Ease Scores*

Score	School level	Definition
100-90	5th grade	Very easy to read. Easily understood by an average 11 year old student
90-80	6th grade	Easy to read. Conversational english for consumers
80-70	7th grade	Fairly easy to read
70-60	8th and 9th grade	Plain english. Easily understood by 13-15 year old students
60-50	10th to 12th grade	Fairly difficult to read
50-30	College	Difficult to read
30-0	College graduate	Very difficult to read. Best understood by college graduates

$$0.39 \left(\frac{\text{total words}}{\text{total sentences}} \right) + 11.8 \left(\frac{\text{total syllables}}{\text{total words}} \right) - 15.59$$

Figure 20: *The Flesch-Kincaid Grade Level formula*

First, a small list of popular websites was formed for targeted investigation. The ToS of these websites were then collected, the version of ToS collected was in effect as of September 5th 2017, and have likely since been updated. The first test was simply to calculate and compare the total word count of these documents.

Next, these ToS were used to investigate their readability. As most websites cater to a large range of user ages, their ToS should reflect this. It is important for users to understand the ToS they agree to. It should then be assumed that ToS should cater to the comprehension level of the users and that the users are actually understanding the content within the ToS document. The readability and comprehension level should be appropriate for the youngest audience that can legally accept the ToS. If users cannot understand (even if they do read) the ToS (or PP) of a website, how can it be assumed that they are giving informed consent? The answer is simple, the user cannot. While this investigation was completed before the implementation of the EU GDPR, it sheds light and discusses some of the issues this regulations aims to rectify.

To test the overall readability of website ToS, two different readability tests have been used. The first, the Flesch readability score measures “reading ease”, and then the Flesch-Kincaid grade score test as the second, measures “grade score”. These initial readability tests were part of an investigation and the results were not directly pursued after the this test. However, these results encouraged the case studies to investigate the legalities and requirements of services in regards to their ToS and PP documents in more depth. These tests also aided the development of concept two, as the output adjusted to address concerns surrounding the users ability to understand the ToS and PP documents.

THE FLESCH READING EASE

In this readability test, the higher score indicates text that is easier to read, while the lower scores indicate text that is more difficult to read. The formula for this test is shown in Figure 19 and the scores are defined in Table 1.

THE FLESCH-KINCAID GRADE SCORE

This readability test indicates the U.S. grade level, intended to make it easier for teachers, parents, librarians, and others to judge the readability of texts (“The Flesch Reading Ease and Flesch-Kincaid Grade Level,” 2017). By indicating the number of years of education the reader needs on average to understand the text. The formula for this test is shown in Figure 20.

The resulting number from the algorithm can be interpreted as the number of years of education the reader requires in order to understand the text. This means it can surpass the limitation of grade levels and still retain significance and understandability.

Table 2: Readability score of popular written work

Text	Flesch Reading Score	Flesch-kincaid Grade Score	Age For Expected Understanding	Total Word Count
Harry Potter And The Philosopher's Stone – J. K. Rowling	82.5: easy to read	4.4: fourth grade	8-9 years old	77846
The Cat In The Hat – Dr. Seuss	112.5: very, very easy to read	-0.7: kindergarten	Toddlers	1621
Twilight – Stephenie Meyer	82: easy to read	5.2: fifth grade	10-11 years old	119901
Looking For Alaska - John Green	84.2: easy to read	4.5: fifth grade	8-9 years old	61770
The Hunger Games (book 1) – Suzanne Collins	86.3: easy to read	4.2: fourth grade	8-9 years old	99750
Pride and Prejudice – Jane Austen	78.1: fairly easy to read	5.5: fifth grade	10-11 years old	122189
Carrie – Stephen King	71.1: fairly easy to read	7.1: Seventh grade	12-13 years old	61326
Romeo and Juliet – Shakespeare	94.4: very easy to read	1.9: first grade	6-7 years old	24545
Great Expectations – Charles Dickens	80.7: easy to read	6.4: sixth grade	11-12 years old	162690
The Adventures Of Huckleberry Finn – Mark Twain	86.6: easy to read	5.6: fifth grade	10-11 years old	109571

Table 3: Readability score of popular websites

Website (.com)	Flesch Reading Score	Flesch-kincaid Grade Score	Age For Expected Understanding	Total Word Count
Google	54.4: fairly difficult to read	10.4: tenth grade	13-14 years old	2033
Youtube	40: difficult to read	13.9: college	18-19 years old	3368
Facebook	42.3: difficult to read	13.2: college	18-19 years old	3518
Wikipedia	42.5: difficult to read	12.9: college	17-18 years old	5857
Reddit	43.7: difficult to read	13.1: college	18-19 years old	3008
Instagram	37.4: difficult to read	14.6: college	19-20 years old	5169
Amazon	41.1: difficult to read	12.8: college	17-18 years old	3377
<i>averages</i>	43.1: difficult to read	13.0: college	18-19 years old	3762

INVESTIGATION RESULTS

To better understand the results, a range of well known texts written for a range of ages, were also passed through the readability scoring algorithms (popular texts are shown in Table 2, and popular websites tested are shown in Table 3). This allowed for comparisons to be made between the ToS and popular texts (such as novels) to help situate the results and give the scores more context for understanding.

The written texts investigated had readability scores that reflected the ages of their respective target reader groups, however this did not hold when applied to the ToS of websites. The closest to reflecting this was Google, which achieved a score that reflects the youngest age group who can legally agree (due to the COPPA regulation). However, most ToS fall under the scores of ‘difficult to read’ or college level complexity, meaning users would only be expected to understand if they were 18 years or older. The only exception to this was Google, who (according to these tests) caters to the youngest audience that can legally accept the ToS themselves (14 year olds). Overall, literature and this exploration shows text should (and can) be written to be appropriate for their target audiences readability skill throughout a range of genres to support the reader (Klare, 2000). And Google showed that ToS documents can also cater to all of their target audience. It can then be assumed, that other online services could cater to the reading levels of all their audience, but are currently neglecting to do so.

While these readability scores are not designed to cater to the ToS documents, this sparked the development of further ideation as the research continued. It is also important to acknowledge the controversy of many readability tests, as many of these tests are not consistently accurate and are often considered arbitrary (CommunicateHealth, 2014; Matters, 2016). The most reliable way to interpret readability scores is to test with multiple and check for correlations, however readability score will always be inherently inaccurate as the measurements will always prove inaccurate in certain use-cases (Giles, 1990).

For the purposes of this research, the accuracy of these readability scores did not prove essential as the result inspired further iterations on the design and a shift away from these metrics. These initial investigations did, however, explain some of the overarching flaws of ToS documents (such as the overarching difficulty in comprehension that was present in the initial context research).

Figure 21: Flowchart depicting the process for developing website ToS summarises

ITERATION TWO: SUMMARY GENERATOR

This iteration was developed alongside this thesis included a system designed to summarise website ToS, ideally in plain english, for internet users to read. This output was designed to be integrated into the users online experience as a browser extension. The concept emerged as a direct result of the concept one and contextual research, which both indicated that users do not read ToS due to length and complicated language. This concept began development as the background research was still being conducted, allowing the background research to directly feed into development.

To understand the process of summarising ToS, before being able to develop the code, manual summary processes were employed. This allowed for extensive investigations into the key themes present, and also the structure or layout of these documents. It was important for the summary process to be comprehensive. Investigations into key themes present was vital to understanding what sections were most important to include in the end summary. These themes were largely directed by the content of the ToS themselves, however, they also included the themes directed by the topics respondents stated they cared most about in the questionnaire.

Exploring how to summarise ToS documents manually also allowed for the solidifying of a process for summarising these ToS. This was then used to create the coded system for automating this process. The process followed a simple order of operations:

- First, scan ToS document for information related to a broad range of topics: data, PII, tracking, 3rd party permissions, policy update notifications, data and/or account deletion;
- After highlighting and taking notes, the relevant sentences were extracted and condensed into plain english (bullet point style);
- Next all the duplicates were found and removed, ensuring only one version existed;
- Lastly, summarised key important points into concise statements.

This system worked effectively for a time consuming process of manual summarisation, however it went under some restructuring when it was shifted into code development (see Figure 21). This was an important shift as it converts a manual process into an automated one. The revised process was as follows:

- Convert the website ToS to plain text (removing all formatting);
- Extract keyword sentences (using the most common five keywords, ignoring stop words);
- Reduce the number of sentences extracted to one sentence per keyword;
- Convert these final selected sentences to plain english.

The first three steps of this process had been completed in development prior to the concept being iterated on and moved towards concept three. However,

the output created during the course of development, while unrefined, was able to begin providing summaries of website ToS.

EXAMPLE SUMMARISES

Google:

MANUAL KEYWORDS:

Keywords: security, privacy, information, personal, control

- By using our Services, you agree that Google can use such data in accordance with our privacy policies.
- If there is a conflict between these terms and the additional terms, the additional terms will control for that conflict.
- By using our Services, you agree that Google can use such data in accordance with our privacy policies.
- Our automated systems analyze your content (including emails) to provide you personally relevant product features, such as customized search results, tailored advertising, and spam and malware detection.
- We provide information to help copyright holders manage their intellectual property online.

AUTOMATED KEYWORDS:

Keywords: services, privacy, terms, additional, content

- Otherwise, all claims arising out of or relating to these terms or the services will be litigated exclusively in the federal or state courts of Santa Clara County, California, USA, and you and Google consent to personal jurisdiction in those courts.
- By using our Services, you agree that Google can use such data in accordance with our privacy policies.
- Our Services are very diverse, so sometimes additional terms or product requirements (including age requirements) may apply.
- Additional terms will be available with the relevant Services, and those additional terms become part of your agreement with us if you use those Services.
- You may not use content from our Services unless you obtain permission from its owner or are otherwise permitted by law.

Facebook:

MANUAL KEYWORDS:

Keywords: security, privacy, information, personal, control

- We also encourage you to review the privacy choices that you have in your settings.
- We do not control or direct what people and others do or say, and

we are not responsible for their actions or conduct (whether online or offline) or any content that they share (including offensive, inappropriate, obscene, unlawful and other objectionable content).

- We use the data that we have to make suggestions for you and others – for example, groups to join, events to attend, Pages to follow or send a message to, shows to watch and people who you may want to become friends with .
- We use the data we have – for example, about the connections you make, the choices and settings you select, and what you share and do on and off our Products – to personalise your experience.
- Permission to use your name, profile picture and information about your actions with ads and sponsored content: You give us permission to use your name and profile picture and information about actions that you have taken on Facebook next to or in connection with ads, offers and other sponsored content that we display across our Products, without any compensation to you.

AUTOMATED KEYWORDS:

Keywords: account, terms, content, apply, services

- We respond to notices of alleged copyright infringement and terminate accounts of repeat infringers according to the process set out in the U.
- Our Services are very diverse, so sometimes additional terms or product requirements (including age requirements) may apply.
- You may not use content from our Services unless you obtain permission from its owner or are otherwise permitted by law.
- Our Services are very diverse, so sometimes additional terms or product requirements (including age requirements) may apply.
- Otherwise, all claims arising out of or relating to these terms or the services will be litigated exclusively in the federal or state courts of Santa Clara County, California, USA, and you and Google consent to personal jurisdiction in those courts.

The generated output was a very crude form of summary, that while they did not lead into any further development, encouraged development into concept three. Concept three included a shift to allow further focus on addressing the factors that influence whether users read ToS and PP.

As research progressed, it became apparent that one of the leading reasons why users do not read these documents is due to the length. While a plain english summary reduced overall length and would, in theory, minimise the readability difficulty, this design solution would only partially solve the issue(s) it set out to. The completion of the contextual research enabled this concept to build the foundation for the third, and final, concept.

ITERATION THREE: RANKING POLICIES WITH MACHINE LEARNING INTEGRATION

The most notable differences between the third concept and the prior concepts, was the shift to Privacy Policy (PP) documents rather than ToS. This aims to directly facilitate developing a solution to the privacy paradox.

After the context research came to a completion, it became apparent that it is important for researchers to begin acting on the conclusions drawn from their research if the privacy paradox is to be mollified. This spurred development of an output that would aim to address many of the current reasons users do not read ToS and PP documents online. The overall scope of this thesis limited the final development time, resulting in this iteration solely focusing on PP (ToS could be included in future).

Initial Concept:

With a combination of considerations into sustainability, functionality, and accessibility, the developed design output changed considerably throughout the ideation process. After considering the sheer number of websites on the internet, it was important to create a design output that was sustainable and had the adaptability to cater to the number of website legal documents that internet users encounter. It was also important to ensure the project had the technology to reflect the scope of what was proposed.

For this to be achieved, the proposed design was a browser extension that would be integrated directly into the users browsing experience. The browser extension is designed with dual functionality: to display concise PP versions to the user, and to highlight PP areas of interest. This aims to encourage further reading of the official PP while also allow specific focus on information, such as security of PII.

The concept was further developed to include a machine learning trained model that ranks predefined categories or themes present in ToS and PP on a scale of 'good', 'average', and 'bad' represented largely through colour and visualisations. The colours used for this scale will be will include a red to green scale (good = green, average = orange, and bad = red). This would provide a visual aid to help users assess how well websites ToS and PP protect user privacy at a glance. This design intended to make website ToS and PP documents as approachable as possible, while also providing minimal clutter and reducing the amount of information presented to the user. This would provide a small amount of information for them then to decide how and if they want to act.

This design aimed to inspire user motivation by providing enough information for them to begin understanding how PII is treated by the websites they visit (in relation to the predefined categories of the browser extension) for them to begin understanding the risks associated with interacting online. This design aimed to relate to Rogers' (1975) user motivation theory and use

the understanding of this theory to help motivate users to take better PII protection behaviours online.

The implementation of ML algorithms allowed for automation of the rank generation process. This enables sustainability, as the ML system will adapt as policies are updated and also minimise the time required for rank generation. This reduces and removes many of the issues (primarily overall sustainability of projects) present in the projects in the Precedent Analysis section.

Due to the limited scope of this thesis, it became apparent during the ideation stage that the limited time for training the machine learning model would limit the accuracy of the proof of concept design of the output. It also became apparent that the ranking system was partially left open for interpretation, despite having an explanation, as different users would classify the ranks differently for the categories than the definitions defined in this research. To help with the identification and explanation of these ranks, the design was adjusted to include a toggleable option that expands out the categories to show the related sentence from the PP document. This aims to both increase the user understanding of the given ranks, and also give the user more information directly from the PP without the requirement to read the whole policy to understand the rank. The option to expand or collapse the extracted sentences also aims to provide a solution to issues of length dissuading users from engaging with website legalese. This gives users the option to engage with the tool with or without lines of text visible.

Refined Concept:

The final concept which moved into development was a browser extension that implements the use of machine learning technology to rate websites on a scale (“good”, “average”, “bad”, with a “n/a” option when more information is required) in specific categories. Previous work and the questionnaire conducted as part of this thesis indicated user topics of interest in PP documents. These topics of interest were then used as the categories for representation in the browser extension. The ranked categories include:

- Data security
- Privacy settings
- Third party tracking
- Policy change alerts
- User control of data
- Data collection
- Account deletion

The option to expand the category to toggle the visibility of the sentences relating to the category was included in the functionality of the extension. This also allows for the user to determine if they agree with the given ranking, especially as the machine learning model does not have a large set of training data. This is also intended to give the user a more curated version of the PP,

Figure 22: Initial aesthetic explorations for the proposed browser extension

specific to topics internet users were proven to be interested in during the questionnaire conducted during the course of this thesis.

This final concept acts as the cohesion point for the iterative designs of the developed outputs and the context written research of the thesis. This combines these two parts and focuses on providing a potential solution to the most significant factors identified that act as the sources of the issues covered in this thesis. This allows for full consideration into all aspects of the design output and the development of a final design that, while not commercially available, aims to address the issues of the research field.

PROJECT DISCLAIMER

This project should not be considered legal advice. This project does not endorse any service in any way and expresses content with no guarantee. The use of machine learning technology to create these outputs does not guarantee that this project is accurate in its representation of the policy documents it is summarising. It is important to note that it is a machine generated representation and not to be treated as a replacement for reading the full ToS and PP documents. Using this project does not in any way substitute for reading the full terms of service and privacy policy which you, as a user, are bound.

Aesthetic Iterations

During this phase, after the final design concept was decided upon, multiple different aesthetics were drafted to represent the information first as an icon or graphic, and then include an intuitive option for the users to expand for more specific information. See Figure 22 for initial iterations, and Figure 23 for iterations of the final aesthetic concept.

SCOPE

The scope of this proposed output can be separated into two different categories: the scope of this thesis, and the future scope of this project.

THESIS DELIVERABLES

The main deliverable to be produced alongside the conducted research, is a proof of concept model of what this project could (with continued development) eventuate into. This proof of concept model will build off the related work addressed in this thesis and be used as a conceptual work to prove the feasibility of a project such as this to help resolve the social issues researched by this thesis (issues such as the privacy paradox).

The output published as part of this thesis is an early-stage version of the proposed output, that can categorise website PP documents for internet users. This project will be a functional browser extension (developed for Google Chrome) illustrating the interaction between the machine learning elements and the integration into the web browser experience. This output will be

Figure 23: Iterations of the chosen aesthetics of the proposed browser extension

used as proof that this project meets both the aims of integrating sustainable technology to ensure design longevity and accessibility for internet users (both in ease of use and making PP documents more approachable for the users to read/interact with).

CONCEPTUAL PROJECT DELIVERABLES

With additional development, this tool would be a fully functional browser extension that seamlessly categorises website ToS and PP. This fully fledged version would include the integration of automated machine learning systems within the extension. This would aim to minimise the required user inputs so the accessibility of both the extension and the legal document(s) is increased to the maximum level possible.

The continued development of this project would include the full proposed functionality, as indicated by the proof of concept, and be released for free. This final version would also include cross-browser compatibility. This would enable a wider user base, and align with the overarching goal of increasing user accessibility. Further development of this project could also include adding additional languages so the project is not restricted to internet users who can read the English language.

Stretch goals for this project, after the full release of the tool, include adapting a version for mobile and tablet users. Allowing for cross device functionality would further the aims of increasing accessibility and further enable this tool to address the privacy paradox and other related issues.

**RE: PRIVACY (PROOF
OF CONCEPT TOOL)**

The final design that was developed has a dual focus functionality to create an early-stage proof of concept example to indicate the proposed design implementation. It can be broken down into two respective parts: the browser extension and the machine learning model.

BROWSER EXTENSION

The browser extension element of the final output displays processed information extracted from PP documents. This extension immediately visualises this information (through the use of colours and icons), with the option for users to extend this information out into more detail through a toggleable expansion of the initial visualization.

This extension focuses on minimising the initial amount of information presented to the user, to help convey information on redefined categories while focusing on removing the dominant reasons users currently neglect to read website PP documents (see Figure 24). This includes reasons such as minimising the overall amount of text, and the time investment to understand the document. By reducing information into visuals, while generalising the overarching content slightly, it enables quicker identification of user respect and rights.

The extension also offers a brief project explanation and includes the disclaimer. This also includes a page explaining the ranking system and categories so the user can actively understand the extension as they engage with it (see Figure 25).

The extension includes a small icon on the top left of the main popup html that indicates if the shown summary has been generated by a human or by the ML model (described in more detail below). This conveys the overall accuracy of the results to the user and aims to help generate trust and transparency of the process to the user interacting with the developed tool.

Figure 24: Default screen of browser extension for a ranked website

Figure 25: Information screen and disclaimer screen within extension explaining project details to user

ML MODEL

The initial proof of concept Machine Learning model built was a collaborative effort between Tom White (a Lecturer at Victoria University of Wellington) and I, that implemented spaCy technology (spaCy, 2018). The spaCy technology is an open source python library for Natural Language Processing (NLP) and was used for the ML model for multiple reasons. These include:

- **Familiarity:** I had previously used the spaCy NLP library. This prior experience with spaCy allowed for faster understanding of the conventions and system used. Working alongside Tom White also allowed for fast prototyping and building of the overall system as he has extensive knowledge in python and machine learning algorithms.
- **Well documented:** SpaCy systems are well documented, allowing for a lower entry level for learning how these models can be implemented for projects and ease of learning how to accurately build the desired outputs.
- **Prebuilt models:** The use of existing technology allowed for a faster design process and it removed focus from recreating existing technology, and instead allowed for sole focus on building the desired output.

The spaCy statistical models allowed for the implementation of a system that generates predictive answers. The statistical model implemented in this project was trained to predict the overall “goodness” of a websites PP (in regard to the data security category). The dataset is currently too small to verify consistent validity of generated predictions (currently the model is trained off twelve pieces of data), however as a proof of concept model this shows how the algorithm could be employed to create the overall functionality of the system. For the future of the project, the training dataset will need to be increased to a number of at least a few hundred. With a larger dataset and further model training this system could be adapted to include the additional intended categories, and would include the ability to accurately test the overall validity of the model.

The system was built using a clearly defined process, designed to create a ML model that is both adaptive and can have a quantifiable accuracy measured. This process can be summarised in Figures 26 (model training) and 27 (model implementation).

The predictive based ML model is trained on the already categorised inputs to determine the overall category ranks of the text (good is given a value of 1 and bad is given a value of 0). The predictive based model generates a number between zero and one, dependent on where it thinks the newly input data is situated on the scale. The generated numbers are based off a combination of factors including the number of model generations (the number of iterations taken to build the model off the data supplied), and the quantity and quality of the data supplied.

The model was trained with twelve categorised website PP documents, focusing on the category “data security”. The model also used four pre-categorised PP

Figure 26: Machine Learning model training process

documents for validity checks during the training process, however as the dataset was too small for the inclusion of these to be significant, this has no overall impact or ability to guarantee the accuracy of the resulting output. The model was trained off a balanced number of “good” and “bad” ranked PP documents, to help ensure the final ML model build was unbiased.

The model underwent two experiments. The first experiment included training the model on only the extracted sentences from the “data security” category. The second experiment was then trained on the full PP document in regards to the selected category.

The model build for the proof of concept design was tested with eighteen website PP documents. These unranked PP were tested with the model that was trained using the full PP document, as opposed to the extracted sentences specific to the category being ranked. This enabled more accurate development of a tool, as it shows the ability to build a tool that can rank for specific categories using the full text as input. Ensuring the trained model could rank specific categories off the entire PP was essential in beginning to establish the overall application of this tool as a completely automated system. Only one category was chosen to be represented in the proof of concept tool—“data security”. The same functionality can be applied to all categories, so it was not essential to generate ranks with the model for other categories.

The ML model outputs can be interpreted by translating the generated numbers to indicate probabilities. Numbers of “> 0.90” are assumed to mean “probably good”, and numbers of “< 0.10” assumed as “probably bad”. This leaves all predictions between as “unknown”. These results are indicated in the table by coloured cells (green = probably good, red = probably bad, no colour = unknown). The generated outputs from these can be seen in Table 4.

Although the manually processed data and intended future of the project aims to include not only these three options, but also an intermediate “average” rating, the model requires future training before this functionality can be included. With a larger dataset, this additional category could be included, however the current corpus used was too small to reliably include this in this version of the developed algorithm. The base functionality of the ML model, however, has proved that the overall proposed functionality should be achievable given more development time and a larger dataset. The base work of the categorising is already been completed in the development of this thesis, and currently to expand to the other categorises all it need is to be trained on the data of additional categorised as well (this was not part of this experiment due to the time restraints).

The additional proposed functionality of the ML aspect of the project, the sentences extraction, was unable to be included in the ML experiments of this thesis as the data sets were too small to reliably build this functionality. However, with continued development, it is realistic to assume this functionality could be included.

Figure 27: Machine Learning model process of use

Table 4: Machine Learning model generated website ranks

Website	ML Generated Rank ("Data Security" category)
rottentomatoes	0.2042954117
plex	0.9533013701
slack	0.9972839355
zotero	0.001014651614
westpac	0.9582341909
steam	0.08750906587
theverge	0.00386054581
TradeMe	0.0639468953
jetstar	0.7901306152
unity	0.9996423721
telegram	0.5517787337
reddit	0.7208216786
qualtrics	0.9704340696
brave	0.6833344102
adobe	0.2706916928
snapchat	0.6370559931
anz	0.001022667275
imdb	0.9835246801

INTEGRATION

As the current version only reached the initial stages of development as a proof of concept model, the two parts of the project currently function independently. However, the extension model has a system in place for handling the input data from the ML model, this process is currently manual only (however, in future this step would be made automatic to remove any required interactions with the programming by the user) and requires the developer to copy across the generated output from the ML model and place this inside the extension code ready for this code to display it in the extension.

DISCUSSION

The privacy paradox has become a societal issue since the rise in popularity of the World Wide Web (Hafner & Lyon, 2001; “The National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign,” 2018). This paradox explains the ambiguous boundary between what is public and private online caused by the prevailing requirement of personal information disclosure to gain access to many services.

PP research typically has been separated into two main categories. First, research that investigates user behaviour online (Hallam, & Zanella, 2016). And secondly, user attention to online legal documents (Hillman, 2005, Wu et al., 2012). The research conducted in this thesis argues that these two aspects of privacy paradox research are not as disconnected as previously suggested. Instead, this thesis claims that these issues are directly linked and also are impacted by user motivation (Rogers, 1995; Youn, 2009; Halam, & Zanella, 2016).

Exploring these aspects together led to identifying one of the cause of the privacy paradox—user disregard of website legal documents (Hallam, & Zanella, 2016; Hillman, 2005). Online services provide ToS and—where necessary—PP for users to educate them on important business activities, such as PII treatment and data security. These legal documents are often presented to the user as lengthy, unformatted text. It has been identified, both in previous research and during the course of this thesis, that this ToS and PP delivery method often leads to users neglecting to engage with these documents, and/or commonly results in users misunderstanding the privacy and security of information online (Dwyer et al., 2007; Youn, 2009).

By identifying this area for improvement the development of a potential solution for the privacy paradox has been proposed. This tool, ‘Re: Privacy’, was developed through an iterative process that developed through considerations into ToS and PP document readability, and targeted elements that impacted users motivation to read document contents. The core elements aimed to be eradicated with this tool were document length, formatting, and time taken to understand the content which were identified by both

previous research (Bakos et al., 2009; Greenberg, 2010; “Do You Read Terms Of Service Contracts?,” 2016; Larose & Rifon, 2007; (Hallam, & Zanella, 2016; Hillman, 2005), and confirmed by the questionnaire conducted as part of this research. This tool proposes a new method of delivering and updating information for internet users from online services.

As privacy is a subjective concept and often varies due to personal preference and culture (Bellman et al., 2004; Wu et al., 2012; Utz & Krämer, 2009), it was important to reflect this in the proposed tool if it is to reflect the scale of the digital environment. By deploying a design directly into the browser experience without changing how users and service providers behave on the platform, it allows this tool to act as a supplement for the users’ online internet experience. As an opt-in design, users are allowed the choice and freedom to engage only if they want. For service providers, Re: Privacy does not change the statements made in their policies, but rather extracts based on topic. This reduces possible misunderstandings of these statements as they are removed from the overall context and setting of the whole policy document. This also provides a faster information relay from provider to user.

Due to the scope of this research, many of the underlying themes were generalised and have not been investigated to their full depth. Given more time, the topics covered would be explored in further detail and the links between them explored to find additional correlations. One of the largest restrictions included the limited scope of the timeline. However, there were also restrictions placed on individual parts of the research itself. The questionnaire conducted as part of the initial process was limited in range and scope of the participant pool. This resulted in a very small local reference. Additionally, this limited the research in its ability to relate findings to a larger scale, and was largely reliant on the validity of previous surveys and questionnaires explored during the initial context research.

Re: Privacy (the proposed design tool) was largely limited by these time restrictions. This was mainly due to the time consuming nature of building datasets and refining the accuracy of the trained ML model. These factors impacted the overall quality of the ML aspect of the tool, as the training dataset was manually created by one researcher. To help counteract the limitations of this, this model largely relied on the work of previous projects discussed in the precedent analysis section, to compile a dataset (these are open source and posted online for other developers to use and review)—the ToS;DR project and the PrivacyGuide project (see the Precedent Analysis section for more information on these projects). In the future, the training dataset would be built to include a larger PP corpus, and more time would be spent refining and building the model to ensure its accuracy. Given more time, the proof of concept tool would be developed to include all defined categories (see Iteration Three included in the Design Output chapter for more information on categories) and be refined into a fully functional product ready for market penetration.

In future, the design proposed in this thesis could be integrated into not only the online browser experience of laptop and desktop users but also it could be developed to include mobile device support. The solution proposed by this research requires user testing to measure adoption and usage to see if it fulfils its aims for self sustainability. However, as an initial proof of concept model is has shown that the correlations between the different research fields investigated allow for a potential privacy paradox to exist. The machine learning technology implemented reinforces and proves that with more development transparency online could be improved to successfully inform users of the impact of sharing PII online.

CONCLUSION

This thesis explores a potential solution to the social phenomenon known as the ‘privacy paradox’ through the exploration of online website legal documents (ToS and PP). While these two research fields (privacy paradox research and legalese research) may at first appear unrelated, some of the leading causes of the privacy paradox can be directly linked to user treatment of online legal documents. Due to factors such as length and complexity, many internet users consistently dismiss reading ToS and PP documents and instead chose to engage with the service(s) without knowing the contents. While some projects in the past have aimed to draw user attention to these documents, few have been successful in highlighting the importance of these documents while also staying updated in the ever changing environment of the internet.

With the growing media attention to this research field due to “data scandals” and the implementation of the EU GDPR, internet users are beginning to become more aware of the treatment of PII online. The pervasiveness of the internet in society has largely encouraged users to actively engage with a multitude of services online, however the impracticality of reading the legal policies of the range of services has only added to the factors discouraging users from actually reading the legal documents.

Users are becoming more aware of the loss of control and protection of their PII on the internet due to many services taking advantage of user neglect to read the legal documents posted. This has created a shift in the privacy paradox, making research into a solution that can pacify all involved parties more relevant in the current social climate. The introduction of new policies and media coverage of large scale data misuse has also encouraged online users to become more conscious of how their online behaviours result in PII disclosure which may not always be as conservative as they initially aim.

The development of a browser extension tool, Re: Privacy, was informed by the contextual research which derived the main factors influencing treatment of these documents, and also the main topics users care most about from the contents. The proposed design integrates ML technology that will penetrate

the browsing experience and also help keep the content of the design relevant as policies are updated. This tool offers the user a curated version of the website PP, represented through clear visualisations to convey information at a glance. This design aims integrates into the browsing experience of the users and supplements the experience, rather than detracting from the overall engagement of the user.

While the overall results of this thesis are not generalised and have yet to undergo usertesting, it provides insight into a development of a potential solution to the privacy paradox. This proposed tool, after more development, could supplement interactions and help bring attention to many concerns users experience online.

REFERENCES

Alpaydin, E., & Bach, F. (2014). *Introduction to Machine Learning*. Cambridge, UNITED STATES: MIT Press. Retrieved from <http://ebookcentral.proquest.com/lib/vuw/detail.action?docID=3339851>

AI-powered Privacy Policies. (n.d.). Retrieved February 26, 2018, from <https://priebot.org/>

Babbie, E. (2013). *The Practice of Social Research*. United States of America: Cengage Learning.

Bakos, Y., Marotta-Wurgler, F., & Trossen, D. (2009). *Does Anyone Read the Fine Print? Consumer Attention to Standard Form Contracts*. Law & Economics Research Paper Series, 09-40.

Barnes, S. B. (2006). *A privacy paradox: Social networking in the United States*. First Monday, 11(9). Retrieved from <http://firstmonday.org/ojs/index.php/fm/article/view/1394>

Bell, J. (2015). *What Is Machine Learning?* In Machine Learning (pp. 01-16). Wiley-Blackwell. <https://doi.org/10.1002/9781119183464.ch1>

Bellman, S., Johnson, E., Kobrin, S., & Lohse, G. (2004). *International Differences in Information Privacy Concerns: A Global Survey of Consumers*. Information Society, 20(5), 313-324. <https://doi.org/10.1080/01972240490507956>

CalOPPA: California Online Privacy Protection Act. (2016, January 5). Retrieved December 4, 2017, from <https://termsfeed.com/blog/caloppa/>

Censorship and the Internet - dia.govt.nz. (n.d.). Retrieved May 30, 2018, from <https://www.dia.govt.nz/Censorship-and-the-Internet>

Censorship Online Safety - dia.govt.nz. (n.d.). Retrieved May 30, 2018, from <https://www.dia.govt.nz/Censorship-Online-Safety>

Clickwrapped - Google. (n.d.). Retrieved February 27, 2018, from <https://www.clickwrapped.com/sites/google>

CommunicateHealth. (2014, June 12). *Rethinking Readability Scores: Part 1*. Retrieved July 11, 2018, from <https://medium.com/wehearthehealthliteracy/rethinking-readability-scores-part-1-703e4e48875d>

Contract and Commercial Law Act 2017 No 5 (as at 14 July 2017), Public Act – New Zealand Legislation. (n.d.). Retrieved May 30, 2018, from <http://www.legislation.govt.nz/act/public/2017/0005/latest/whole.html#DLM6844033>

COPPA - Children's Online Privacy Protection. (n.d.). Retrieved November 27, 2017, from <http://www.coppa.org/>

Davies, H. (2015, December 11). *Ted Cruz campaign using firm that harvested data on millions of unwitting Facebook users*. The Guardian. Retrieved from <http://www.theguardian.com/us-news/2015/dec/11/senator-ted-cruz-president-campaign-facebook-user-data>

Do You Read Terms Of Service Contracts? Not Many Do, Research Shows. (2016). Retrieved May 28, 2018, from <https://www.npr.org/2016/08/23/491024846/do-you-read-terms-of-service-contracts-not-many-do-research-shows>

Dwyer, C., Hiltz, S., & Passerini, K. (2007). *Trust and privacy concern within social networking sites: A comparison of Facebook and MySpace*. AMCIS 2007 Proceedings, 339.

E&C Announces Facebook CEO Mark Zuckerberg Will Testify April 11th. (2018, April 4). Retrieved May 29, 2018, from <https://energycommerce.house.gov/news/press-release/ec-announces-facebook-ceo-mark-zuckerberg->

will-testify-april-11th/

EP Conference of Presidents with Mark ZUCKERBERG - founder and CEO of Facebook| European Parliament. (2018). Retrieved May 29, 2018, from <http://www.europarl.europa.eu/ep-live/en/other-events/video?event=20180522-1820-SPECIAL-UNKN>

European Commission. (2016). Directive 95/46/EC (General Data Protection Regulation). Retrieved from: <http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679>

Exposed: Undercover secrets of Trump's data firm. (2018, March 20). Retrieved May 29, 2018, from <https://www.channel4.com/news/exposed-undercover-secrets-of-donald-trump-data-firm-cambridge-analytica>

Facebook, Social Media Privacy, and the Use and Abuse of Data | United States Senate Committee on the Judiciary. (2018). Retrieved May 29, 2018, from <https://www.judiciary.senate.gov/meetings/facebook-social-media-privacy-and-the-use-and-abuse-of-data>

Federal Trade Commission Act. (2013, July 19). Retrieved July 13, 2018, from <https://www.ftc.gov/enforcement/statutes/federal-trade-commission-act>

Gatepeepin'. (n.d.). Retrieved February 28, 2018, from <http://www.gatepeepin.org/>

Giles, T. D. (1990). *Readability Formulas and Technical Writing*. Baywood Publishing Co., Inc.

Greenberg, A. (2010). *Who Reads The Fine Print Online? Less Than One Person In 1000*. Retrieved May 28, 2018, from <https://www.forbes.com/sites/firewall/2010/04/08/who-reads-the-fine-print-online-less-than-one-person-in-1000/>

Hafner, K., & Lyon, M. (2001). *Where wizards stay up late: the origins of the Internet*. New York: Touchstone. Retrieved from <https://www.overdrive.com/search?q=49B402D8-7A9A-4462-8E25-D3F99425B1B2>

Hallam, C., & Zanella, G. (2016). *Online self-disclosure: The privacy paradox explained as a temporally discounted balance between concerns and rewards*. *Computers in Human Behavior*, 68, 217–227. <https://doi.org/10.1016/j.chb.2016.11.033>

Hanington, B., & Martin, B. (2012). *Universal Methods of Design*. United States of America: Rockport Publishers.

Hern, A. (2018, May 25). *Facebook and Google targeted as first GDPR complaints filed*. *The Guardian*. Retrieved from <http://www.theguardian.com/technology/2018/may/25/facebook-google-gdpr-complaints-eu-consumer-rights>

Hillman, R. A. (2005, March 15). *Online Consumer Standard-Form Contracting Practices: A Survey and Discussion of Legal Implications*. Cornell Law Faculty Publications.

Hofmann, P., Nakamura, T., Kiyomoto, S., Serna, J., & Tesfay. (2018). *I Read but Don't Agree: Privacy Policy Benchmarking using Machine Learning and the EU GDPR* (pp. 163–166). ACM Press. <https://doi.org/10.1145/3184558.3186969>

Jordan, M. I., & Mitchell, T. M. (2015). *Machine learning: Trends, perspectives, and prospects*. *Science*, 349(6245), 255–260. <https://doi.org/10.1126/science.aaa8415>

Klare, G. R. (2000). *The measurement of readability: useful information for communicators*. *ACM Journal of Computer Documentation*, 24(3), 107–121. <https://doi.org/10.1145/344599.344630>

Krippendorff, K. (2003). *Content Analysis: An Introduction to Its Methodolgy* (2nd Edition). Sage Publications, Inc.

Lanzing, M. (2016). *Pasts, presents, and prophecies on your life story and the (re) collection and future use of your data*. In *The Art of Ethics in the Information Age* (pp. 83–87). Amsterdam: Amsterdam University Press.

Larose, R., & Rifon, N. J. (2007). *Promoting i-safety: Effects of privacy warnings and privacy seals on risk assessment and online privacy behavior*. *Journal of Consumer*

Affairs, 41(1), 127–149. <https://doi.org/http://dx.doi.org/10.1111/j.1745-6606.2006.00071.x>

Liao, C., Liu, C.-C., & Chen, K. (2011). *Examining the impact of privacy, trust and risk perceptions beyond monetary transactions: An integrated model*. *Electronic Commerce Research and Applications*, 10(6), 702–715. <https://doi.org/10.1016/j.elerap.2011.07.003>

Linda Hilfling Gate peepin | NIMk. (2008). Retrieved July 25, 2017, from <http://www.nimk.nl/eng/linda-hilfling-gate-peepin>

Matters, H. W. in C. (2016, November 24). *Flesh of my Flesch: Readability Matters · Prism Clarity*. Retrieved July 11, 2018, from <https://prism-clarity.com/2016/11/flesh-kincaid-readability-matters/>

Mayer, J. R., & Mitchell, J. C. (2012). *Third-Party Web Tracking: Policy and Technology*. In 2012 IEEE Symposium on Security and Privacy (pp. 413–427). <https://doi.org/10.1109/SP.2012.47>

noyb.eu | My Privacy is none of your business. (2018). Retrieved May 27, 2018, from <https://noyb.eu>

Obar, J., & Oeldorf-Hirsch, A. (2016). *The Biggest Lie on the Internet: Ignoring the privacy policies and terms of service policies of social networking services*. https://www.ftc.gov/system/files/documents/public_comments/2016/10/00067-129185.pdf

O'Donnell, K., & Cramer, H. (2015). *People's Perceptions of Personalized Ads*. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 1293–1298). New York, NY, USA: ACM. <https://doi.org/10.1145/2740908.2742003>

Orwell, George. (1949). *Nineteen Eighty-Four*. London: Penguin Group.

Park, Y. J., Campbell, S. W., & Kwak, N. (2012). *Affect, cognition and reward: Predictors of privacy protection online*. *Computers in Human Behavior*, 28(3), 1019–1027. <https://doi.org/10.1016/j.chb.2012.01.004>

Privacy, simplified. — DuckDuckGo Browser Extension & Mobile App.

(n.d.). Retrieved February 28, 2018, from <https://duckduckgo.com/app>

Roderick, L. (2014). *Discipline and Power in the Digital Age: The Case of the US Consumer Data Broker Industry*. *Critical Sociology*, 40(5), 729–746. <https://doi.org/10.1177/0896920513501350>

Rogers, R. W. (1975). *A Protection Motivation Theory of Fear Appeals and Attitude Change*. *Journal of Psychology*, 91(1), 93.

Roszak, T. (1986). *The Cult of Information: The folklore of computers and the true art of thinking*. New York, NY, USA: Pantheon Books.

Samuel, A. L. (1959). *Some studies in machine learning using the game of Checkers*. *Ibm Journal of Research and Development*, 71–105.

spaCy · Industrial-strength Natural Language Processing in Python. (2018). Retrieved May 15, 2018, from <https://spacy.io/>

Suchman, L. (2012). *Configuration*. In Lury, C., & Wakeford, N. (Ed.). *Inventive methods: The happening of the social*. London: Routledge.

Terms of Service; Didn't Read. (n.d.). Retrieved August 14, 2017, from <https://tosdr.org/>

The Cookie Law Explained. (n.d.). Retrieved March 19, 2018, from <https://www.cookielaw.org/the-cookie-law/>

The Flesch Reading Ease and Flesch-Kincaid Grade Level. (2017, January 26). Retrieved July 3, 2018, from <https://readable.io/blog/the-flesch-reading-ease-and-flesch-kincaid-grade-level/>

The National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign. (n.d.). Retrieved July 11, 2018, from <http://www.ncsa.illinois.edu/>

Utz, S., & Krämer, N. C. (2009). *The privacy paradox on social network sites revisited: The role of individual characteristics and group norms*. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 3(2). Retrieved from <https://>

cyberpsychology.eu/article/view/4223

Wu, K.-W., Huang, S. Y., Yen, D. C., & Popova, I. (2012). *The effect of online privacy policy on consumer privacy concern and trust*. *Computers in Human Behavior*, 28(3), 889–897. <https://doi.org/10.1016/j.chb.2011.12.008>

Youn, S. (2009). *Determinants of Online Privacy Concern and Its Influence on Privacy Protection Behaviors Among Young Adolescents*. *Journal of Consumer Affairs*, 43(3), 389–418. <https://doi.org/10.1111/j.1745-6606.2009.01146.x>

LIST OF FIGURES

Unless otherwise stated, all figures were created by the author.

FIGURE 1:

Terms of Service; Didn't Read. (n.d.). Retrieved February 27, 2018, from <https://tosdr.org/>

FIGURE 2:

Terms of Service; Didn't Read. (n.d.). Retrieved February 27, 2018, from <https://tosdr.org/#google>

FIGURE 3:

Privacy, simplified. — DuckDuckGo Browser Extension & Mobile App. (n.d.). Retrieved February 27, 2018, from <https://duckduckgo.com/app>

FIGURE 4:

Polisis. (n.d.). Retrieved February 27, 2018, from <https://pribot.org/polisis>

FIGURE 5:

Pribot: The Privacy Bot. (n.d.). Retrieved February 27, 2018, from <http://pribot.org/bot>

FIGURE 6:

Linda Hilfling Gate peepin | NIMk. (n.d.). Retrieved July 25, 2017, from <http://www.nimk.nl/eng/linda-hilfling-gate-peepin>

FIGURE 7:

Clickwrapped. (n.d.). Retrieved February 27, 2018, from <https://www.clickwrapped.com/>

FIGURE 8:

Clickwrapped - Google. (n.d.). Retrieved February 27, 2018, from <https://www.clickwrapped.com/sites/google>

FIGURE 9:

Tesfay, W. B., Hofmann, P., Nakamura, T., Kiyomoto, S., & Serna, J. (2018). I Read but Don'T Agree: Privacy Policy Benchmarking Using Machine Learning and the EU GDPR. In Companion Proceedings of the The Web Conference 2018 (pp. 163–166). Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee. <https://doi.org/10.1145/3184558.3186969>

APPENDIX

Terms of Service Repurposed

INFORMATION FOR PARTICIPANTS

You are invited to take part in this research. Please read this information before deciding whether or not to take part. If you decide to participate, thank you. If you decide not to participate, thank you for considering this request.

Who am I?

My name is *Elizabeth Calvert* and I am a Masters student in DSDN 593 at Victoria University of Wellington. This research project is work towards my thesis.

What is the aim of the project?

This project investigates online user response to website Terms of Service and Privacy Policy documents and how users interact with fingerprint on the internet. The aim is to discover how users treat online Terms of Service and Privacy Policies, and how engagement with these documents can be improved. This research has been approved by the Victoria University of Wellington Human Ethics Committee approval number: 0000025314.

How can you help?

You have been invited to participate because this research aims to encompass a diverse range of internet active members of society. If you agree to take part you will complete a survey. You will be asked questions about your experience and behaviours on the internet, with a focus placed on your opinion on internet website Terms of Service and Privacy Policy documents. The questionnaire will take approximately 5-10 minutes to complete.

What will happen to the information you give?

This research is anonymous. This means that nobody, including the researchers will be aware of your identity. By answering it, you are giving consent for us to use your responses in this research. Your answers will remain completely anonymous and unidentifiable. Once you submit the questionnaire, it will be impossible to retract your answer. Please do not include any personally identifiable information in your responses.

What will the project produce?

The information from my research will be used in my Masters thesis. This information will also be used to aid in the development of a design output alongside my thesis by providing

key insights of possible user expectations and preferences of this design. This design output will be a prototype of a design that may be developed further after the completion of my thesis. The final thesis and research portfolio will be published online as a website. Potentially this will also be included in a journal article or conference paper.

If you have any questions or problems, who can you contact?

If you have any questions, either now or in the future, please feel free to contact either:

Student:

Name: Elizabeth Calvert

Email address: calverliz@myvuw.ac.nz

Supervisor:

Name: Walter Langelaar

Role: Supervisor

School: School of Design

Phone: 04 4635512

Email: Walter.Langelaar@vuw.ac.nz

Human Ethics Committee information

If you have any concerns about the ethical conduct of the research you may contact the Victoria University HEC Convenor: Associate Professor Susan Corbett. Email susan.corbett@vuw.ac.nz or telephone +64-4-463 5480.

Questionnaire Questions:

1) What is your gender?

- ☐ Male
- ☐ Female
- ☐ Other
- ☐ Prefer not to answer

2) What is your age?

- ☐ 0 - 18*
- ☐ 18 - 24
- ☐ 25 - 34
- ☐ 35 - 45
- ☐ 45 - 54
- ☐ 55 - 64
- ☐ 64 - 74
- ☐ 75 +

** If they answer 0 - 18, the survey will link them to the end thanking them for their time.*

3) How many hours per week do you spend browsing the internet?

- ☐ None
- ☐ Less than 5
- ☐ 5 - 10
- ☐ 11 - 15
- ☐ 16 - 25
- ☐ 26 - 35
- ☐ 36 +

4) What platform(s) do you use to access the internet? (Tick all that apply)

- ☐ Laptop/Desktop computer
- ☐ Phone
- ☐ Tablet
- ☐ Other (please specify) _____

5) What services do you actively engage with online? (Tick all that apply)

- ☐ Entertainment (e.g. movies, gaming, music)
- ☐ Communication (e.g. email, instant messaging)
- ☐ Online clubs/groups
- ☐ Competitions/contests
- ☐ Shopping
- ☐ Education
- ☐ Research/information gathering
- ☐ Finance (e.g. online banking)
- ☐ Current events (e.g. news, weather, concerts)
- ☐ Travel related (e.g. reservations)
- ☐ Other (please specify)

6) What type of personal information are you comfortable with sharing on the internet? (Tick all that apply)

- ☐ Name
- ☐ Relationship status
- ☐ Personal address
- ☐ Email address(es)
- ☐ Date of birth/Age
- ☐ Credit card details
- ☐ Medical information
- ☐ Location
- ☐ Phone number(s)
- ☐ Family details
- ☐ Personal photographs
- ☐ Other (please specify)

7) What is your preferred web browser?

- ☐ Google Chrome
- ☐ Firefox
- ☐ Safari
- ☐ Internet Explorer
- ☐ Other (please specify) _____

8) How often do you install browser add-ons to your preferred web browser?

- ☐ Regularly
- ☐ Often
- ☐ Sometimes
- ☐ Seldom
- ☐ Never

9) Which website do you visit most frequently?

- ☐ Google
- ☐ Facebook
- ☐ Reddit
- ☐ Instagram
- ☐ Youtube
- ☐ Other (please specify) _____

10) How concerned for your privacy are you while you are using the internet?

- ☐ Extremely
- ☐ Very
- ☐ Moderately
- ☐ Slightly
- ☐ Not at all

11) Have you ever been concerned that any of the following could occur while you are engaging with services online? (tick those that apply)

- ☐ Identity theft
- ☐ Financial loss
- ☐ Personal information could be misused
- ☐ Personal information could be used by others without your knowledge
- ☐ You will receive spam/junk emails
- ☐ Experiences of discomfort while interacting online

12) What actions do you take to protect privacy online?

13) How likely are you to identify trustworthy sites online?

- ☐ Always
- ☐ Often
- ☐ Sometimes
- ☐ Seldom
- ☐ Never

14) How do you determine if a website is trustworthy?

15) How is personal information disclosed by users online treated by companies?

16) To what extent is personally identifiable information accessible to companies once it has been uploaded online?

17) How often do you read website Terms of Service?

- ☐ Always
- ☐ Often
- ☐ Sometimes
- ☐ Seldom
- ☐ Never

18) How often do you read website privacy policy documents?

- ☐ Always
- ☐ Often
- ☐ Sometimes
- ☐ Seldom
- ☐ Never

19) In your own words, please explain what website "terms of service" are:

20) In your own words, please explain what a website's "privacy policy" is:

21) How important do you believe it is for users to read website terms of service and privacy policy before agreeing to the conditions?

- ☐ Extremely important
- ☐ Very important
- ☐ Moderately important
- ☐ Slightly important
- ☐ Not at all important

22) Which do you believe is more important to read and understand: website terms of service or website privacy policies?

- ☐ Terms of service
- ☐ Privacy policy
- ☐ Both
- ☐ Neither
- ☐ Other (please specify) _____

23) What are the factors that influence whether you read website Terms of Service?

24) What is the key information in Terms of Service you consider most important?

25) What is the key information in Privacy policies you consider most important?

26) Have you ever refused to agree with website Terms of Service? (Please explain your answer)

- ☐ Yes
- ☐ No

27) How often do you compare the Terms of Service/privacy policies of online services before selecting which service to use?

- ☐ Always
- ☐ Often
- ☐ Sometimes
- ☐ Seldom
- ☐ Never

28) When receiving notifications of updated Terms of Service, do you read the updated version?
(Please explain your answer)

- ☐ Yes
- ☐ No

29) If there was an option to read a simplified version of website Terms of Service, how likely do you believe you would read this version?

- ☐ Extremely
- ☐ Very
- ☐ Moderately
- ☐ Slightly
- ☐ Not at all

Thank you for your participation in this survey.