ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Quantifying Mechanisms behind Cookie Consent (Non-)Compliance: A Notification Study of Audit Tools

Bachelor Thesis

Laura-Vanessa Soldner

August 13, 2023

Advisors: Karel Kubíček, Dr. Amit Zac, Ahmed Bouhoula, Prof. Dr. David Basin

Department of Computer Science, ETH Zürich

Abstract

The rise of Big Data companies and the AI Boom has made data ownership and usage a central concern. This shift prompted an in-depth exploration of the data economy's impact, resulting in the creation of privacy laws like the General Data Protection Regulation (GDPR). However, numerous studies reveal that websites struggle with full compliance. Our study aims to quantify the mechanisms of cookie consent compliance and assess the impact of privacy notifications on websites. We propose three mechanisms for non-compliance – legal gaps, technical gaps, and translation gaps. We use an audit crawler to find non-GDPR-compliant cookies on over 3000 websites.

We notify those websites regarding their non-compliance using three different types of email notifications. Each notification type focuses on one of the proposed mechanisms. To test legal gaps we provide websites with a report based on the findings of the audit crawler. To quantify technical gaps we inform websites about the (free) self-audit tool *CookieAudit*. To understand the translation gaps we provide websites with both the report and CookieAudit.

We provide descriptive statistics of remediation based on daily crawls before and after notifications. Initial findings show a high fluctuation in the number of observed violations across crawls, obscuring the potentially subtle remediation rate due to our notifications by the noise. However, as the experiment is ongoing, a conclusive analysis is not feasible yet, so we propose analysis options addressing the high data variance. Our study contributes insights into theory-driven notification studies and addresses challenges with inconsistent crawlers.

Acknowledgements

I want to express my thanks to Karel Kubíček and Amit Zac for giving me this opportunity and for their invaluable guidance throughout the entire process. Karel, your patience in answering my questions and sharing your insights has helped me acquire new skills and grasp complex concepts. Amit, your honest feedback and challenging questions at all times of the day have been a driving force every step of the way.

I would also like to extend my thanks to David Basin for allowing me to pursue this thesis in his group. I am truly grateful to Ahmed Bouhoula, Saskia Wolf, the ID-Helpdesk team, and the ISGINF-Helpdesk team for generously sharing their expertise and supporting this thesis. Lastly, special thanks go to Luc for his unwavering support and understanding throughout this endeavor and proof reading the thesis.

Contents

Contents **iii**

1 Introduction **1**

2 Background **4**

2.1 Literature 4

2.1.1 Compliance Research 4

2.1.2 Privacy Notification Research 6

2.2 Mechanisms and Hypotheses 7

2.3 Contributions 10

3 Experiment Design **12**

3.1 Treatments 12

3.2 Experiment 14

3.3 Scope of this thesis 16

3.4 Pre-Registration 16

3.5 Research Ethics 16

4 Collecting Non-Compliant Websites **18**

4.1 Website Sampling 18

4.2 Presence and Consent Crawler 19

4.2.1 Adjustments 19

4.2.2 Results 19

4.2.3 Limitations 20

4.3 Database Processing 21

4.4 Email Sampling 21

4.4.1 Email Crawler: 22

4.4.2 Hunter.io: 22

4.4.3 Manual Search: 23

4.4.4 Duplicates 23

4.4.5 Limitations 23
4.5 Assignment of Treatment 24

5 Sending Email Notifications **25**
5.1 CookieAudit 25
5.1.1 Adjustments 26
5.2 Report 26
5.2.1 Design 27
5.2.2 Violation Types 27
5.3 Study Website 27
5.4 Mailbox 27
5.5 Sending Emails 28

6 Results **29**
6.1 Descriptive Statistics 29
6.1.1 Independent Variable and Control Variables 29
6.1.2 Reliability of the Crawler 30
6.1.3 Dependent Variable: Crawl Results 32
6.2 Responses to the Notifications 35
6.3 Discussion 36

7 Conclusion **41**

Bibliography **43**

A Email Texts **47**
A.1 First Contact Email 47
A.1.1 Treatment Self-Check 47
A.1.2 Treatment Report 49
A.1.3 Treatment Self-Check and Report 51
A.2 Reminder Email 55
A.3 Debriefing and Survey Email 56
A.3.1 Control Group 56

B Study Website **59**

C Survey **61**

Chapter 1

Introduction

Data has undoubtedly become the oil of this era. While for a long time considered a natural byproduct of internet usage the ownership and rightful monetization of data was not of general societal interest. The rise of Big Data companies like Meta or Alphabet and the AI Boom in the past years however have made the topic of data indispensable in our daily lives. It forced academia and civil society to look in depth at the impact of the data economy and to address legal and social issues. In Europe, more than in other parts of the world, privacy is considered a fundamental right [1, Art. 8] establishing a framework of legislation aimed at protecting online privacy.

The legal framework includes EU-wide level legislation such as the *ePrivacy Directive* [2], enforced since June 31st, 2002, and the *General Data Protection Regulation (GDPR)* [1], enforced since May 25th, 2018, which serve as the legal ground for analyses in this thesis. The ePrivacy Directive requires that any non-essential data collection and processing requires consent, and the GDPR requires this consent to be freely-given, informed, active, and unambiguous. Users also have the right to withdraw consent at any time [1, Art. 4, 7]. Further, users have the right to access their data, rectify inaccuracies, request erasure, and obtain their data [1, Art. 15, 16, 17, 20].

The GDPRs scope extends beyond the EU, applying to entities and websites outside the EU that process the personal information of EU residents, ensuring comprehensive protection for EU citizens' data regardless of processing location [1, Art. 3]. Ensuring compliance with the consent requirements can be legally challenging for website operators. As a result, many choose to utilize a third-party solution known as a *Consent Management Provider (CMP)*. The CMP takes care of the legal terms and conditions for privacy policies, implements the consent banner, collects user consent, and manages its redistribution [3]. However, using a CMP does not guarantee automatically full compliance with privacy law and does not change the legal liability of a website owner towards its users.

Previous research has shown that many websites violate these obligations [4, 5, 6]. Only a few qualitative, small-scale studies have explored potential reasons for non-compliance [7, 8, 9]. Based on their findings, we synthesize three main mechanisms. First, the *the Legal Gap* highlights insufficient understanding of the legal requirements of the privacy framework. Second, *the Technical Gap* consists of significant knowledge gaps concerning technical implementations of the requirements. Third, *the Translation Gap* is the specific interaction between Legal and Technical Gaps and highlights issues in translating legal policies into technical ones.

This thesis contributes to this issue on three levels. First, we explore the gap theories based on an empiric and theory-driven approach. The aim is to better understand the concerns and needs of website operators. Second, we study the effect of a notification about a self-help tool. This not only allows testing of the Technical Gap but also contributes a new element to notification research. Third, we further contribute to privacy notification research by analyzing the effect of different framings on notifications about complex privacy issues regarding cookie consent.

We achieve this by designing a randomized control field experiment. Using a modified crawler by Bollinger et al. [4], we automatically scan websites monitoring their GDPR violations of cookie requirements. We configure the crawler to visit websites in Germany, Ireland, and France to analyze their cookie consent compliance. Based on this we create a study sample of websites that violate laws on cookie consent. Those websites receive different types of email notifications with each focusing on one gap theory. After creating the sample we run the crawler daily on this sample to track the compliance of websites before and after the treatment. As the experiment is still running when completing the thesis the next steps are outlined but not yet carried out. We send a reminder to the websites after one month, and after another month we send a debriefing email with a survey.

After the completion of the study, we will analyze the crawl data to measure remediation rates. Remediation rates refers to the percentage of identified violations that have been successfully fixed. The survey should provide insight into the motivations and challenges of websites. Lastly, we will discuss email responses we received from the website during the study to better understand the effect of our notification.

Our observations reveal significant variability in the crawler's daily website visits and the cookies it identifies in each run. Factors like connectivity disruptions and server load fluctuations may contribute to this variance. The variance of flagged cookies depends upon the type of the cookie violation.

As of now, we have not observed a reduction in the number of flagged cookies among the treatment groups in comparison to the control group, regardless of violation types. This outcome is not unexpected, considering our notifications address intricate cookie issues and likely result in low remediation rates. Moreover, our analysis includes only the initial two weeks post-treatment, which may not provide websites enough time to address these concerns. The impact of privacy notifications on GDPR compliance remains inconclusive based on the existing data. Importantly, comprehensive details, including notification emails, survey responses, and full two-month crawler data, are still pending. As such, definitive conclusions cannot yet be made.

In Chapter 2, we discuss the theoretical background of why a majority of websites are not GDPR compliant and formulate concrete hypotheses to test. In Chapter 3, we describe the design of the experiment. Following, the process of adjusting the crawler, creating the website sample, and gathering the emails is described in Chapter 4. Chapter 5 outlines further steps we took before notifying the websites related to the notifications and the emails. Chapter 6 provide a descriptive analysis of the data up to two weeks after we set the treatment. This chapter also includes a discussion of potential limitations and the preliminary results at the time this bachelor thesis was written. Finally, Chapter 7 concludes the thesis.

Chapter 2

Background

In the following section, the relevant literature is summed up, structured, and discussed. The gap in the literature is identified and the importance of filling it is illustrated. We lay out the theoretical basis of the compliance mechanisms and present three specific hypotheses for testing. Finally, we highlight our contribution.

2.1 Literature

The study of GDPR[1] compliance is complex. The GDPR itself is extensive and covers a wide range of privacy rules. Assessing compliance is challenging because there is no straightforward method to verify whether a website follows all privacy laws. Despite this, a majority of studies find that many websites violate at least one requirement of the GDPR [4, 5, 10, 6].

As this thesis contributes to two areas of research, the literature is split into two parts. First, we explore research on how to measure GDPR compliance and the mechanisms fostering compliance. Second, we summarize previous research on privacy notifications.

2.1.1 Compliance Research

Several publications right after the enactment of the GDPR in May 2018 found that while there is a measurable effect on tracking, user-tracking is not reduced long-term [11, 12, 13]. Those early publications do not specifically differentiate between compliant and non-compliant behavior.

[1] While many publications focus on the GDPR, it is important to note that some aspects discussed in this section and henceforth are obligations from the ePrivacy Directive. Therefore, in this thesis, when referring to the GDPR, the obligations outlined in the ePrivacy Directive are also meant to be included. Both sets of obligations are commonly referred to as privacy laws.

Measuring Non-Compliance Studies that attempt to measure GDPR compliance usually focus on a subset of regulations. There have been several studies conducted on different subset areas. Concerning consent banners, Nouwens et al. [14] suggested only 11.8% of observed websites were fully GDPR compliant. Utz et al. [10] found 57.4% of observed websites use dark patterns to prone users into giving consent and Matte et al. [6] found violations regarding cookie consent and consent banners in 54% of the observed websites. Bollinger et al. [4] analyzed 30,000 websites for eight cookie violations such as ignoring user consent, wrong or missing cookie labels, and incorrect retention periods. They found that 94.7% of the analyzed websites show at least one potential violation. More recent studies support these findings [5, 15]. The focus of all of those publications remains on developing technical instruments to measure compliance and estimating how many websites are not fully GDPR compliant. The underlying mechanism producing these patterns remains unclear.

Translating the GDPR from Legal to Tech There are publications that suggest specific actions websites can take to become GDPR compliant. Ayala-Rivera et al. [16] and Mohan et al. [17] offer concrete approaches to guide website owners. They implicitly assume that non-compliant websites lack the know-how to transform legal text into concrete technological policies. However, the reasons behind non-compliance remain vastly unexplored.

Understanding the Mechanisms A handful of qualitative studies took upon the question of what challenges organizations face in implementing the GDPR. Sirur et al. [7] conducted interviews with 12 participants before the GDPR came into force. They highlight that for large companies and companies with an affinity for security, the implementation of the privacy laws was not a big challenge and the organizations were in general positive about the change in the law. For small- and medium-sized organizations without affinity for security, the implementation of the GDPR posed bigger challenges. Specifically translating the legal text into concrete technical policies, a lack of internal resources, and missing guidance from national data protection intuitions was highlighted. For some organizations becoming GDPR compliant was not a priority. Freitas and da Silva [8] interviewed representatives of 12 small- and medium-sized companies. They primarily find that those companies struggle with legal know-how on how to approach the GDPR but mostly lack awareness about the privacy laws altogether. Li et al. [9] followed 8 employees for several months and documented their implementation of the GDPR. They find three driving mechanisms for non-compliance: missing awareness, a lack of tools to check compliance, and missing incentives because of the company's business model. All of those studies are exploitative, small-scale, and qualitative. Theory-driven research on compliance mechanisms is sparse and to the best of our knowledge, no quantitative approach has yet been used to further back these findings. This thesis measures the effect of two of those mechanisms: missing awareness and lack of tools to check compliance.

2.1.2 Privacy Notification Research

In the context of privacy issues, it is established for researchers to resort to the means of notifying companies directly. Experiments on how effective notifications can be designed have been conducted even before the GDPR [18, 19]. There are several publications on notifications on GDPR compliance. In comparison to this thesis, these notification studies build upon existing notification theory and do not use notifications as a means to better understand compliance mechanisms. They only provide information that the issue exists but do not actively offer information on how to remediate the problems.

Maass et al. [20] compared the effectiveness of different contact media, different senders, and different framings and found that letters from a legal research group are the most effective. They measured remediation rates nearly 50 percentage points higher than compared to no notification. The issue they notify websites about is a wrong Google Analytics setting. This concerns non-compliance with a very small and specific subset of privacy laws. Further, this issue can be fixed within 5 minutes to an hour, depending on the know-how of the employee. It remains open how these findings translate to more complex privacy issues. They also survey the notified websites on their experience with the notification and the GDPR in general.

Utz et al. [21] scale previous studies to a larger sample size. They use a high level of automation for detecting non-compliant websites, gathering email addresses, and sending out notifications. In total, they contact over 115,000 websites. As a treatment, they compare the effect of security notifications to privacy notifications. They identify three possible privacy issues: no privacy policy on the website, no cookie notice on the website, and the use of HTTP for sensitive information. The issues were detected with automated tools. Manual validation yielded that some of the issues have false positive rates of up to 6.8%. For the security issue, they check if the source code of a website is accidentally publicly accessible. They find that privacy issues have a lower remediation rate. Overall, remediation rates are very low, for most issues less than 1% with no issue above 3%. This might be a result of many websites not receiving the notification and that these issues are more difficult to fix compared to the Google Analytics issue chosen by Maass et al.[20]. This study of Utz et al. provides valuable insights into the challenges of scaling privacy notifications. In terms of difficulty to fix the issues, we conclude from this to expect similar remediation rates. We assume that fixing those types of issues takes from around an hour to a day depending on how familiar the employee is with the setup of the website. They also include a survey that shows that most websites perceive the notifications positively.

Compared to the other studies described above, Henning et al. [22] focus on better understanding the motivations behind (non-)compliance. They looked at issues with the design of cookie banners, specifically if an opt-out option is available and if colors are used to nudge a user. They notified 147 websites about the violations and send a survey to better understand the reasons for the design implicitly trying to understand compliance behavior. The remediation rate was up to 8%. We assume that fixing the problem takes less than a day and is more straightforward compared to Utz et al. and our study. Seven websites responded to the survey which makes it difficult to find general mechanisms. The participants for the most part did not see a problem with their design, i.e., with non-compliance. The response rates and remediating rates were higher for a friendly non-threatening framing. Given the small sample size, it is unclear whether their findings can be generalized.

2.2 Mechanisms and Hypotheses

To theorize the mechanisms behind non-compliance with privacy laws, we summarize and structure the existing research on this topic and then suggest hypotheses for testing.

Based on previous research on compliance and privacy notifications three main drivers of non-compliance can be differentiated. First, Sirur et al. [7] and Li et al. [9] highlight that some companies lack the know-how or resources to implement privacy obligations. Further, they might be limited in their pursuit of compliance because they cannot check their website for compliance. This suggests a gap in technical know-how. Like in the case of the legal gap discussed next, this is a plausible assumption as many websites rely on Consent Management Providers (CMPs) to implement cookie consent notices. Hils et al. have shown that since the enforcement of the GDPR, the usage of CMPs has quadrupled, which strongly supports this assumption [23]. Organizations prefer to outsource the service than develop an in-house solution. The complexity and size of the GDPR also makes it hard to ensure that each article has been addressed. Therefore, our definition of the first potential mechanism:

Mechanism 1. *Technical Gap: A website understands the legal requirements on a heuristic level but does not know what implementations are needed to fulfill those. A website fails to formulate concrete technical policies. A website cannot test or check its implementations.*

Sirur et al. [7], Freitas and da Silva [8], and Li et al. [9] all found that some companies lack awareness or know-how on how to tackle the GDPR from the legal perspective. Some small- and medium-size companies state that they are not sure or were not aware that the privacy laws apply to them in the first place. This indicates a gap in legal know-how. This is a plausible assumption as the GDPR is a comprehensive framework and small and medium-sized companies do not necessarily have employees with a legal background. Again the raise of CMPs demonstrates this. Thus, we define the second potential mechanism as follows:

Mechanism 2. *Legal Gap: A website is not aware of specific GDPR articles or the GDPR as a whole. A website might be aware of the laws but does not know how to interpret them and formulate them into company-wide privacy policies. A website might rely on a national data protection office for guidance but does not receive such in the desired amount.*

We believe that the two mechanisms are somewhat intertwined. Translating legal obligations into technical implementations [7, 9, 8] seems to be a challenge of its own. This suggests some overlap between the technical gap and a legal gap may exist, but also an interaction effect between them as one issue exacerbates the other. It could also mean that improvement in compliance requires closing both gaps simultaneously. Thus, we define the third mechanism:

Mechanism 3. *Translation Gap: A website shows elements of a legal gap and a technical gap. Additionally, it struggles to translate legal requirements into technical implementation policies.*

In addition to the two main mechanisms, we want to discuss two other factors which might influence compliance behavior.

First, Freitas and da Silva suggest that a reason companies are not compliant is the lack of (human) resources [8] to implement the changes in time. However, their study was conducted shortly after the enforcement of the GDPR. As it is now over 5 years after the enactment we argue that even companies with fewer resources – in the sense of human resources and time – have had the time to become compliant which is why this mechanism is not further explored here.

Second, Sirur et al. [7] and Henning et al. [22] found that for some companies the GDPR is not of interest or not a priority. This aligns with findings from Henning et al. [22]: Despite being notified about an issue (and thus a potential fine) some website owners actively decide against adjusting their website. High levels of compliance might also conflict with the business model of a company behind a website. Overall, these reactions indicate benefits of non-compliance might be higher than the potential drawbacks and costs. This is plausible as detection and prosecution rates of GDPR violations are low [24] and websites might consider non-compliance a low-risk. In addition, it seems that most enforcement activities are aimed at large (data) companies like Google and Meta with some national exceptions. This might suggest lower incentives for small- and medium-sized organizations. Websites behave rationally, calculating the risk of fines and enforcement compared to immediate benefits. However, this theory does not suggest an isolated mechanism, but a broader perspective on compliance in general. Becker's *crime and punishment model* [25], and following developments, suggest that the decision to comply with the laws are affected by the scope of laws, size of fines (punishment), and perceived detection rates compared to benefits from the violating behavior. It relates to the two mechanisms described above, as closing the legal or technical gap changes the risk associated with the decision. In other words, while the legal context is constant, a change in the legal, technical, or translation gaps of a website might change their perception of the context. We address the issue, as much as possible, by making sure confidentiality is guaranteed, i.e., websites do not feel as they are further exposed to enforcement activities.

We will test the mechanisms with three different email notifications. The first notification offers free access to a self-check tool and aims to fill Technical Gaps. The second notification shares a compliance audit report with information on each compliance issue with the website. This should reduce Legal Gaps. A third notification includes both the self-check tool and the audit report and aims to mitigate issues of translation Gaps. The notifications are explained in more detail in Chapter 3.

Based on these mechanisms and the three treatments we derive three hypotheses. First, we want to understand if there are gaps of any type in the first place. The literature on notification research generally finds a positive effect of notifications on remediation [20, 21, 22] suggesting that websites exhibit knowledge gaps. Thus our first hypothesis:

Hypothesis 1. *Any treatment will improve the compliance of a website compared to the control (no treatment).*

We can test this thesis by comparing the compliance of websites receiving any notification compared to websites not receiving any notifications. Second, we want to better understand the interaction effect suggested by the Translation Gap. Thus, our second hypothesis:

Hypothesis 2. *A treatment combining information to close both Legal and Technical Gaps improves compliance of a website more compared to each treatment separately.*

We test this by comparing the effects of websites receiving a notification with both the self-check tool and the audit report in comparison to websites receiving information on only one of them. Given the small sample size of previous studies, it is not clear which mechanism affects compliance the most. The legal gap was mentioned more often in the qualitative studes [8, 9, 7]. Further, there are paid services that provide audit reports similar to ours. This suggests that websites are willing to pay to close this gap. We suggest that the Legal Gap is more critical and the notification including the audit report will have a larger effect on compliance than the self-check tool. Thus our third hypothesis:

Hypothesis 3. *A treatment providing information to close Legal Gaps improves the compliance of a website more than a treatment providing information on how to close the Technical Gap.*

This hypothesis is tested by comparing the effect size on the increase in compliance of the *Self-Check* treatment and the *Report* treatment. All of the treatments might have an effect on the perceived risk of non-compliance. A website might have assessed the benefits of non-compliance higher than the risks before our treatment but might reevaluate this assessment after our notification. However, these effects are difficult to measure, especially as we do not have data on how many website representatives actually read our email actively. However, we attempt to measure the effect to some degree with our survey.

2.3 Contributions

Advancements in researching technical tools to measure non-compliance and estimating non-compliance rates regarding the GDPR have inevitably led to the question of why compliance rates are so low. So far, only qualitative studies have been conducted. There has been no large-scale quantitative study to address this question. Understanding the mechanisms, motivations, and incentive structures of website owners is crucial. By doing so, privacy campaigns can be designed more effectively and the responsible law enforcement bodies can be relieved. The overall compliance could be increased by better understanding the needs of the companies. When contacting companies, we can gain better insights into the effectiveness of notifications by adopting a theory-driven approach.

This thesis also contributes to privacy notification research. Firstly, we propose a paradigm shift by basing the notification design on compliance theory. Furthermore, no study has yet conducted international, large-scale notification experiments with different framings for complex privacy issues, specifically problems with cookie consent. This is especially concerning as publications suggest that non-compliance for cookie consent is over 90% [4]. This thesis aims to analyze the effectiveness of such notifications. Additionally, previous research suggests that websites would appreciate to receive more information on how to fix the addressed issue. Compared to other notification study our goal is further to actively empower websites and test the effectiveness notifications in combination with self-audit tools.

Chapter 3

Experiment Design

In this chapter, we describe the treatment groups designed to answer hypotheses from the previous chapter, then we document the data collection of our experiment, and finally, we refer to the pre-registration of this study.

3.1 Treatments

As outlined in the previous chapter, our treatment consists of three different email notifications. Each website receives one type of notification except for the control group which does not receive any notification. Each of the three notifications is aimed to test a specific mechanism.

The first type of treatment consists of notifying a website about a self-check tool. The tool we suggest is named *CookieAudit* and was developed at the Information Security Group at ETH Zurich [26]. It allows users to audit websites, including their own, detect potential privacy law violations, share flagged cookies, and provides suggestions on how to rectify the violations. The tool can be downloaded from the Chrome Web Store and is free to use. In the notification, we share the link to CookieAudit in the Chrome Web Store. The results of this tool are not collected by the developers. In conclusion, the first treatment is defined as follows:

Treatment Group 1. *Self-Check: The website receives an email from an ETH-designated email introducing ourselves as a legal-computer science research group. We offer free access to a self-check tool to test the compliance of the website via a link to the Chrome Web Store. We do not inform the website of any potential violations we have found.*

The primary objective of analyzing the remediation rate in response to the *Self-Check* treatment is to comprehensively understand the impact of the Technical Gap. The *Self-Check* treatment offers a mechanism for website operators to enhance their adherence to privacy regulations and data protection standards. By employing CookieAudit, websites can proactively assess their compliance status, identify flagged cookies of concern, and access detailed guidance on resolving any identified issues.

The second type of treatment consists of a notification including an audit report of the website. This report includes information about types of violations and lists specific cookies we have found in this category. The audit report is based on the audit crawler developed by Bollinger et al. [4]. Their crawler is an automated instrument used for analyzing large numbers of websites. The audit report is appended as a PDF to the notification email. The second treatment is defined as follows:

Treatment Group 2. *Report: The website receives an email from an ETH-designated email introducing ourselves as a legal-computer science research group. We share the results of our audit tool in the form of a personalized PDF report attached to the email. We do not inform the website about the self-check tool.*

The remediation rate in response to the Report treatment provides insights into the Legal Gap. Organizations might struggle to understand the provisions of the GDPR. To address this, the Report treatment is designed to furnish comprehensive information on cookie consent issues, accompanied by the relevant GDPR Articles. By doing so, it provides organizations with the essential knowledge to effectively tackle potential privacy compliance challenges. Incorporating specific GDPR articles in the Report treatment serves a dual purpose. Firstly, it enhances comprehension of the legal framework governing data protection and privacy rights, enabling organizations to better grasp their responsibilities and obligations under the GDPR. Secondly, by linking identified cookie consent issues directly to corresponding legal provisions, the Report raises awareness of GDPR obligations.

The third type of treatment consists of a notification including both a link to the self-check tool and the audit report. It is a combination of the two notifications described above. Specifically, we define it as:

Treatment Group 3. *Self-Check AND Report: The website receives an email from an ETH-designated email introducing ourselves as a legal-computer science research group. We share the results of our audit tool in the form of a personalized report attached to the email and we offer free access to a self-check tool to test the compliance of the website.*

This combined treatment offers a comprehensive perspective on the accumulative effect of the Technical Gap and the Legal Gap in privacy compliance behavior. It incorporates elements from both the *Report* and *Self-Check* approaches. This combination enhances comprehension of the legal framework, reinforcing their understanding of GDPR responsibilities. Additionally, the *Self-Check* mechanism, facilitated by CookieAudit, enables operators to proactively assess compliance status, identify flagged cookies, and access detailed guidance for rectification, effectively addressing the Technical Gap.

Lastly, we have the Control Group, which does not receive a notification email:

Treatment Group 4. *Control Group: No contact at all. We review their websites at all time points.*

The exact wording of the email notifications can be found in Appendix A. The treatments link to the hypotheses in the following way: For the first hypothesis we compare remediation rates between websites that received a notification (*Self-Check*, *Report*, or *Self-Check AND Report*) and websites that did not receive a notification (control group). For the second hypothesis, we compare remediation rates between websites that receive the notification *Self-Check AND Report* compared to websites that receive a notification on only one of the tools (*Self-Check* or *Report*). Finally, for the third hypothesis, we test if the remediation rate is higher for websites that receive a notification with a report compared to websites that receive a notification linking the self-audit tool.

3.2 Experiment

This section explains how our experiment is set up. Our experiment follows the structure of previous notification experiments conducted by Maass et al. [20] and Utz et al. [21]. It can be divided into several steps:

Collecting Non-Compliant Websites: To create a dataset of websites with potential violations, [1] we use the automated web crawler developed by Bollinger et al. [4]. The crawler also provides a list of specific cookies that may be causing potential violations. As input, we select websites accessible within the EU to fall under the GDPR. The input sample is described in more detail in Chapter 4. After the crawl, approximately 3,000 websites remain, which we contact via email.

To gather email addresses to contact the website operators, we employ several approaches: an automated crawler, the services of an email finder company,[2] and manual searches on the websites. To protect the privacy of

[1] The automated crawler by Bollinger et al. [4] and the self-check tool CookieAudit [26] were developed with the best intentions and provide robust and reliable information. However, they may occasionally produce false positives. To avoid any unintended consequences, in all communication with the websites such as the notification emails we refer to these as *potential* issues and violations. For readability, this distinction is usually omitted in this thesis.

[2] https://hunter.io/

website employees, we only collect generic info email addresses from the email finder company. We manually validated 100 of the email addresses obtained through the first method to ensure the accuracy of the email crawler. Emphasizing the validity of the email addresses reflects our commitment to high ethical standards and aims to prevent any unwanted negative responses. We used email notifications despite Mass et al.'s [20] observation that letters are a more effective notification medium. Scaling up letter notifications can be challenging, which reinforces the value of better understanding email notifications themselves. This step is further explained in Chapter 4.

Sending Email Notifications: After obtaining the sample of websites violating GDPR, we randomly assign them to one of four groups: treatment *Self-Check*, treatment *Report*, treatment *Self-Check AND Report*, and the control group. Websites in the treatment groups will receive an email notification concerning privacy law compliance. It is crucial to ensure that website representatives comprehend our notifications clearly, without any misunderstandings. To prevent potential side effects, we focus on websites in Germany, Ireland, and France, where we are proficient in the corresponding languages. For websites in Germany and France, we send emails in both English and the respective local language (German or French).

Previous publications have shown that reminders can positively impact remediation [20, 21]. Hence, one month after the initial notification, we send a reminder email.

After another month, a debriefing email is sent to the websites, including an invitation to participate in a survey. Although response rates for surveys in this context are generally low [20, 21], even a small number of responses can provide valuable insights. The purpose of the survey is to offer websites the opportunity to provide more information about their motivations and needs regarding privacy law compliance (see Appendix C). Additionally, we aim to gather more data on the organization's size, industry, and business model. This step is further explained in Chapter 5.

Runtime: The websites are crawled daily starting from a week before the treatment, during the treatment, and for some weeks after the treatment. The runtime of the experiment of two months has been determined to be enough time for organizations with few resources to have the chance to become compliant. After two months, the data from the cookie crawler will be used to obtain information on the number of cookie consent violations and thus on the (non-)compliance. This data and the responses to the survey will be analyzed. Furthermore, any messages and inquiries from organizations will be encoded and will also be discussed.

3.3 Scope of this thesis

At the time of writing and completing this thesis, the experiment is still ongoing. As a result, not all steps described above are included here. This thesis covers the process of designing the study, selecting the website sample, gathering the emails, and sending out the first notification. The result section will discuss the remediation behavior observed during the first two weeks after the treatment. The notification email will not be sent out yet. The experiment is intended to continue to its completion, and if possible, will be published in an academic journal.

3.4 Pre-Registration

A pre-registration is a transparent process where researchers publicly register their research plan before data collection. It enhances research rigor, transparency, and reproducibility by committing researchers to specific methodologies and analyses, reducing questionable research practices. It promotes openness, accountability, and credibility in scientific inquiry.

The design of this study was pre-registered in the Open Science Framework, which is maintained by the Center for Open Science.[3] We decided to use the pre-registration template from AsPredicted.org. This is a shorter pre-registration which we determined a better fit for this interdisciplinary study.

3.5 Research Ethics

We sought approval from the ETH Ethics Commission for this study to ensure the ethical integrity of our research. Our approach was designed to avoid collecting any personal data beyond what is publicly accessible, such as email addresses from websites or generic identifiers. Adhering to the principle of minimal data usage, we will anonymize survey responses to prevent the identification of specific websites or email addresses used for notifications. Our study design was approved by the ETH Ethics Commission with minor modifications.

We do not employ any deception in the study. Still, we will offer websites an informal "debriefing" in the email which notifies them about the survey and informs them in more detail of the study's aims (see debriefing email in Appendix A and survey in Appendix C). After our communication with the websites, when the information is no longer needed for processing purposes, we will anonymize the dataset. This also applies to the crawled data. Websites are assigned unique IDs with no relation to the website. No personal information (like name, specific age, etc.) of the representative

[3] The permanent link of the pre-registration is https://doi.org/10.17605/OSF.IO/BYPKR

will ever be collected during the study. All data is completely anonymized in accordance with ETH Law Art. 36d. The results, including the crawl data and the survey responses, are published as aggregates of anonymized information, with no connection to individual websites.

We estimate the risk to be minimal, while we offer an important contribution to both science and society as policymakers continue to struggle to protect users' privacy and to enforce websites' compliance. An effort that has mostly proven unfruitful. We offer a way forward by working with website owners while answering impactful policy questions.

Chapter 4

Collecting Non-Compliant Websites

This section provides information about the sampling process and the audit crawler. It further explains how the raw data from the crawler was processed and how the email addresses were sampled. Finally, we describe how the treatment was assigned to the final sample.

4.1 Website Sampling

Previous publications used various website rankings, among others lists by Tranco, TheInternetBackup project, and Wikipedia [4, 21, 20]. Bollinger et al. [4] used the Tranco list. Recent research on the correctness and consistency of these lists suggests that the Chrome User Experience Report (*CrUX*) currently provides the most accurate data [27]. Furthermore, due to the discontinuation of the Alexa list, which partially serves as the basis for the Tranco list, starting from February 1st, 2023, we opted to utilize CrUX instead [28, 27].

The CrUX dataset is collected by Google and includes measurements of the performance of websites and web applications. CrUX gathers anonymized data from users who opt into the *Chrome User Experience Report* service. The dataset contains a diverse range of performance-related metrics that can be broadly categorized into two main groups: loading metrics and interactivity metrics. Loading metrics measure the speed at which a website's content becomes visible to users. Interactivity metrics evaluate how responsive a website is to user interactions. It further includes an aggregated ranking of the website popularity. The data is structured in monthly batches. It can be freely accessed via BigQuery or the CrUX API, we opted for the latter.

In the first step, we sample the 100,000 most popular websites in the countries Germany, Ireland, and France from the CrUX dataset from the month of April 2023. A website can be among the most popular 100,000 websites in several countries. As every website should be in the dataset only once, we assign it to the country in which it is most popular. This reduces the final set to 248,277 websites.

4.2 Presence and Consent Crawler

To get the website compliance analysis we use the crawler developed by Dino Bollinger [4]. The first part of this crawler is the presence crawl. The presence crawl prepares the set of websites for the consent crawl. Given the original set of websites, it removes websites to which it could not establish a connection and websites that do not use a CMP from either OneTrust or Cookiebot. The consent crawler visits each of the present websites and randomly navigates the site by scrolling and clicking on subpages iteratively. After a fixed, limited number of clicks, the crawler analyzes cookie consent compliance of eight different categories. It returns a dataset with non-compliant cookies which were detected per violation per website. Such cookies are referred to as *flagged cookies*.

4.2.1 Adjustments

The main adjustment was to set up an automated pipeline to run the consent crawl every day for a longer period of time. The crawler was dockerized which allows for more consistent runs and a higher level of abstraction and automation. We adjusted the crawler such that by setting a Boolean variable the crawl either includes the presence crawl or leaves out this step and starts the consent crawl. The input set of websites for the consent crawl can be adjusted by the user. For the study, we always use the websites set for the consent crawl. We set up a pipeline for daily runs which creates a new output folder for every run. The raw databases, a log on the crawl, a log on which input data was used (in the case that we only run the consent crawl), and the final dataset are saved.

4.2.2 Results

We run this crawler dockerized and using a VPN endpoint located in Germany. To generate the sample we run the presence crawl once with the CrUX data. After the run 4237 websites using Cookiebot and 5594 websites using OneTrust remain in the set. This means that all other websites either are not reachable with the crawler or do not use Cookiebot or Onetrust. Based on this first sample we run the consent crawl five times. A website is included in the next sample if it has data for at least one of those five runs. This sample contains 3780 websites.

To gather information on those websites over time, we use the CRON library. We automatically start the consent crawler with the fixed set of present websites daily at 00.15 am. A CRON script also creates automatic backups every day. The crawler ran daily from June 27th to July 6th. Due to human error, the crawler did not run from July 7th to July 11th. It then ran again daily from July 12th on. At the moment of writing this thesis (August 13th), the crawler is still running daily and is expected to run until around September 25th, 2023 – two weeks after the debriefing email. Due to the non-determinism, the crawler typically runs between 10 to 16 hours every day. The crawler outputs SQLite Databases in the range of 0.5 to 1 GB per day.

4.2.3 Limitations

This section discusses in more detail some of the issues and limitations we faced with the audit crawler.

OpenWPM The crawler by Bollinger et al. [4] is based on the OpenWPM library [29], which browses the website with the Mozilla Firefox browser. Their crawler uses version *v*0.12.0 of OpenWPM, which was developed in 2020.[1] OpenWPM did a significant rewrite in 2021 which broke backward compatibility and in general struggles with supporting older code versions. Within the scope of this thesis, it was not possible to rewrite the crawler such that it works on a newer OpenWPM version. However, using the older *v*0.12.0 version of OpenWPM in the crawler posed another issue. It fails to automatically install the correct Mozilla Firefox version.[2] We manually downloaded and included Mozilla Firefox version 80.0.1. Running the crawl with an outdated browser impacts our dataset significantly.

First of all, the presence crawler detects fewer websites as many websites are not accessible via outdated browsers. Second, this biases our set of websites towards websites with a lower security standard. A website with high investments in security will not be reachable over an outdated browser. This might also shift our dataset towards small- and medium-sized organizations as they tend to have fewer resources and know-how to implement such changes. The impact of this on our results will be discussed in Chapter 6.

Stabilization Stabilizing a crawler is difficult. The amount of data per day might vary depending on which web pages were randomly browsed, connectivity issues, changes in URL structures across websites, or fluctuations in server load. Further, the duration of the crawl changes daily. If for example a crawl runs more than 24 hours or if the server runs out of memory, the next crawl does not start automatically. Thus, not all the crawls were started at the same time and due to the non-determinism of the crawl itself, the output

[1] https://github.com/openwpm/OpenWPM/releases/tag/v0.12.0

[2] https://github.com/openwpm/OpenWPM/issues/964

data it produces varies every day. We will account for this in Chapter 6. This is not a unique problem, Utz et al. for example also experienced fluctuations in the crawler output [21].

CMPs Finally, the crawler can only analyze websites using certain CMPs. Thus, our data only includes websites that use the services of Cookiebot or OneTrust. This limits our dataset as neither of those CMPs dominates the market and represents only a fraction of a countries most popular websites. It is not clear whether and how this biases the results. Further, it also excludes websites that do not use an external service to implement cookie consent obligating but use their own solution. This further biases our website set by excluding websites that likely have high resources and know-how in IT.

4.3 Database Processing

The dataset returned by the crawler is unstructured. We took the following steps before sending out the emails. The data that the crawler returns only contains flagged cookies, the type of violation, and the URL of the website. We group the violations for each URL and merge this data with the CrUX dataset to add the country. We merge this with data from the presence crawl to get the CMP for each website. In this thesis, we focus on five types of cookie consent violations from the eight types that the consent crawler produces data for. The reason behind this is that the self-check tool CookieAudit only looks at a subset of the eight violations detected by the crawler. This step removes 777 instances from our dataset, leaving 3003 websites. While this is a substantial reduction we argue that this is necessary for the success of the treatment *Self-Check AND Report*. Unequal information in the two tools can lead to unnecessary confusion and mistrust about the correctness of our notification. Also, we would provide more information to close the Legal Gap compared to the Technical Gap: If a website knows about more violations than it can fix with the self-check tool CookieAudit, we leave them in an unfortunate position making them knowingly but unwillingly non-compliant.

4.4 Email Sampling

After creating the final sample we have 3003 unique URLs in our dataset which represent our email sample. To make our approach scalable we determined to automate as much of the email sampling process as possible. We used three different approaches to sample email addresses which are described below. Our goal was to gather privacy and data protection-related email addresses or generic company email addresses (for example `privacy@example.com`). We refrained from contacting individual employees.

4.4.1 Email Crawler:

We used an automated crawler developed by the Information Security Group at ETH Zurich which browses each website and gathers all email addresses it can find listed on the websites. The crawler found 2609 out of the 3003 website's email addresses. To evaluate the quality of the crawled addresses, we validated 100 data instances manually. We ensured that the email addresses were actually on the website. We also checked if any relevant email address (such as privacy-related email addresses) were missing. Eight out of the hundred manually checked websites included an email address which we did not find on the website. For five websites, we found a relevant email address on the website which was not included in the email-crawl. To select one final email address from the list of addresses that the crawler returned for each website, we used the the heuristic described in Algorithm 1.

Algorithm 1 Heuristic for Choosing Email Address for a Website

```
procedure ChooseWebsiteLanguage(list of emails)
    if email contains ['gdpr', 'rgpd', 'dsgvo', 'dpo'] then
        choose first email in the list containing the keyword
    else if email contains ['privacy', 'daten', 'data', 'donnees', 'données'] then
        choose first email in the list containing the keyword
    else if email contains ['info'] then
        choose first email in the list containing the keyword
    else if email contains ['contact'] then
        choose first email in the list containing the keyword
    else if email contains ['support', 'complaint'] then
        choose first email in the list containing the keyword
    else if email contains the domain of the URL [url_domain] then
        choose first email in the list containing the keyword
    else
        choose the first email in the list
    end if
end procedure
```

4.4.2 Hunter.io:

As 394 websites were still missing, we decided to use a commercial email-finding service called *Hunter.io*[3]. They provide the option to access only generic company email addresses for the selected URLs which we opted for. With this service, we were able to gain additional 167 email addresses.

[3] https://hunter.io/ (last access August 13th 2023)

If several email addresses were outputted for a website, we used the same heuristic as described in the heuristic above (Algorithm 1).

4.4.3 Manual Search:

Finally, 227 websites had no email address yet. We gathered those addresses manually. If several email addresses were displayed on the website, we used the same heuristic as described above (Algorithm 1) to select the final email address.

4.4.4 Duplicates

Even though some false generic email addresses were filtered out in the heuristic, we manually deleted more email addresses after gathering them all. We went over all websites which had an email address assigned that appeared more than once in the dataset and checked the validity of this address. This affected 267 websites. There were two main reasons for duplicates:

Use of same external service: In some instances, the email crawler included email addresses from websites that provided a third-party service to a website. For example, when including a Google Forms on a website, the crawler also analyzed the corresponding Google website linked under the Google Forms. Thus, for all websites which used an email address from a big third-party service provider like Google or Zoom, we manually removed those email addresses and manually searched for the correct ones.

Same Company: In some instances, the duplication of the email address was not a mistake but correct behavior. In this case that one email address is responsible for several websites that have cookie consent issues we wanted them to receive multiple notifications for each website which is why those duplicates have been left in the dataset. However, we ensured that all notifications were of the same treatment type to ensure consistency and avoid confusion for the website owner. In a few instances (less than five) there occurred a duplication of email addresses which has no apparent reason. Those instances were corrected manually. Given the validity checks for the email crawler, we do not suspect structural inconsistencies or problems with the crawler and are positive about the validity of the email addresses despite those few irregularities.

4.4.5 Limitations

The types of email addresses range from privacy related to support and contact email addresses, which might influence how the email is handled within the organization. Further, we decided not to include contact forms from websites. While some websites might prefer being contacted via their contact form (and thus are more likely to read our notification) we argue that using contact forms is not only a lot of manual work but also does not guarantee that the treatment is applied equally as contact forms differ in the type of information that needs to be provided.

4.5 Assignment of Treatment

We randomly assign one of the four treatment groups to each website. As explained above we constrain the assignment to ensure that URLs with the same final email address receive the same treatment. We also ensure in the treatment assignment that there is an equal distribution of treatments within each of the three countries France, Germany, and Ireland. The final treatment group size are shown in Table 6.1.

Chapter 5

Sending Email Notifications

This section discusses the steps that we took to successfully send out the different treatment notifications. A lot of focus went into presenting our notifications and treatments as authentic and trustworthy as possible.

5.1 CookieAudit

CookieAudit is a browser extension developed at the Information Security Group at ETH Zurich [26]. It lets users self-audit websites. The extension detects potential violations of privacy laws, displays flagged cookies, and suggests approaches on how to fix the violations. It differentiates between four different violations:

1. *Non-essential Cookies:* The GDPR and ePrivacy Directive specify that consent must be given explicitly for any use of collection or processing of information which is not strictly necessary [1, 2]. These cookies are flagged if they are declared as not strictly necessary but were collected even when not all cookie purposes were allowed by the website user.

2. *Undeclared Cookies:* Articles 1, 7, and Recital 32 of the GDPR state that consent must be freely given, specific, informed, and unambiguous indication of the user's agreement to the processing of personal data relating to him or her [1, Art. 1, 7, Rec. 32]. Cookies are flagged as undeclared if they are observed being set inside the browser but are not listed in a website's consent notice. This violates the GDPR as first, the user is not informed about those cookies (no informed consent). Second, the user has no possibility to reject those cookies (no freely given consent).

3. *Wrongly Categorized:* Articles 1, 7, and Recital 32 of the GDPR states that consent must be specific and informed [1, Art. 1, 7, Rec. 32]. We flag cookies as wrongly categorized if their classification differs from

what is declared in the consent notice. For instance, 'Google Analytics' cookies cannot be classified as strictly necessary for the operation of the site. If they are wrongly categorized the user does not have the option for (correctly) informed consent. If they are categorized as strictly necessary the user has no possibility to reject those cookies (no freely given consent).

4. *Wrong Expiration Time:* Article 13 of the GDPR [1, Art. 13] and Article 29 of the Working Party (29WP) [30], an EU body that advises on the interpretation of the EU cookie directive, define the necessary information that needs to be declared regarding cookies. This information includes the storage period of the cookie. We flag cookies in two cases: (1) when the actual expiry of a cookie is more than 1.5 times the declared period, and (2) when cookies are declared as session cookies but are actually persistent, or vice versa.

5.1.1 Adjustments

A few adjustments to the first version of CookieAudit were made to make the tool as attractive and helpful as possible to the website owners. We adjusted the layout to make the browser extension more readable. We also adjusted the information boxes for each of the four violations. It now includes more information on the type of violation and contains hyperlinks to the corresponding privacy laws. We also added an introduction text on the purpose of CookieAudit, the research team behind it, and a disclaimer that the results are not deterministic. All those changes were published as version 0.3 of CookieAudit.[1]

5.2 Report

In order to share the findings of the audit tool we decided to create a PDF which we attach to the notification email. We decided that a PDF might be more trustworthy than a link and we did not have to worry about hosting the reports on a server. Our goal was to design the report as similar as possible to the CookieAudit tool in order to have coherency and appear more trustworthy and professional to the websites. All reports were created and sent in English language independent of the associated country to the website.

[1] https://chrome.google.com/webstore/detail/cookieaudit/hoheefgkoickpgelfgijnjnifcpkmbnc (last accessed August 13th, 2023)

CookieAudit Categories	Audit Crawler Categories
Non-Essential Cookies	Ignored User Consent, Implicit Consent
Undeclared Cookies	Undeclared
Wrongly Categorized	Wrong Label
Wrong Expiration Time	Inconsistent Expiry

Table 5.1: Description of how the categories from the Audit Crawler were matched to the CookieAudit categories.

5.2.1 Design

To keep the design as similar as possible, we took the HTML code from CookieAudit as a starting base. We added the ETH logo for more trustworthiness. The text fragments describing the violations in the report are more detailed compared to CookieAudit and contain references to court judgments. However, the content is very similar. We also excluded information on the CMP, URL where the cookie was found and the type of cookie on the report. This is mainly due to technical reasons.

5.2.2 Violation Types

As mentioned before, the audit crawler differentiates between eight categories. We use five of those to match the four categories of CookieAudit as shown in Table 5.1

5.3 Study Website

To ensure the websites that our email notification is trustworthy and we are an actual research team, we added a sub-page on the Information Security Group ETH website about our study. This additional step of ensuring organizations about the authenticity of our study has been used before in similar studies [20, 21]. The exact text of the website can be found in Appendix B

5.4 Mailbox

In order to send out the email we created a separate mailbox from which the notifications were sent and the reminder and debriefing will be sent. That way all researchers have access and we can use a study-specific email address. We use a university-specific address to ensure trustworthiness – not only for the reader but also for the spam filter. While the email address itself is not tied to a specific person, we adjusted the mailbox such that it displays the name of a researcher as the sender.

5.5 Sending Emails

In preparation for sending out the first email notifications, we informed the professor and administrative personnel in our group and the ETH Telefonzentrale about our study and briefed them on how to respond in case an inquiry asking about the authenticity of the study came up. Further, ETH temporarily increased the maximum number of emails that can be sent per 24 hours for us. To send out the emails the Add-on MAILMERGE in Mozilla Firefox was used.[2] It allows to personalize email messages and attachments and review the emails in the outbox before actually sending them. For the notification, nine different types of emails were sent out: three different treatment notifications in three different languages (all including the English version). The notification emails were sent out on July 17 2023 between 07.30 am and 08.40 am. To mitigate the potential of human mistakes in this process, a sending log was created and used during sending.

[2] https://addons.thunderbird.net/en-US/thunderbird/addon/mail-merge/ (last accessed August 13th, 2023)

Chapter 6

Results

This section describes the data for this experiment. As the experiment is still running we reduce this part to descriptive statistics and make concrete suggestions for further analysis. The first section provides an overview of the relevant variables and reviews the reliability of the crawler. It provides correlation plots for a first analysis of the potential treatment effect. In the second section, we discuss the findings and suggest next steps.

6.1 Descriptive Statistics

The following section analyzes the sample described in Chapter 4 by discussing the range of the variables and their distribution in the sample. This section focuses on understanding the behavior of the crawler and potential side effects which shall be accounted for in later work.

6.1.1 Independent Variable and Control Variables

This section gives a brief overview of the data for the independent variable (the treatment group) and the control variables. First, all websites are removed if they have less than one successful crawl per week before the treatment. A successful crawl means that the crawler reached the website and was able to conduct the analysis. This removes 658 websites out of 3003. We had invalid email addresses for 15 websites. As they did not receive a treatment they will be excluded from the analysis. The number of websites that received each treatment, itemized per country, are displayed in Table 6.1. In total, 2330 websites are included in this sample. As we accounted for the countries when assigning the treatments the treatment size groups overall and per country are similar in size.[1]

[1] The reason it is not exactly the same is because of how we handled multiple websites sharing the same contact email address. Our goal was to notify the responsible party about

Country: Overall we have the least websites from France and the most in Germany in our sample (see Table 6.1). This is because there are more OneTrust and Cookiebot users in Germany and Ireland than in France. We include this control variable to account for potential national differences. For example, the impact of the national data protection agency could vary per country. Table 6.3 shows the average number of found cookies before the treatment per country. We can see that overall websites offering services in Germany use fewer cookies than France and Ireland. Interestingly, this is in line with findings we made while gathering emails: German websites often wrote the @ symbol in their website email address as *"at"* or (at) on their websites to prevent automated crawlers from harvesting them. However, those observations were made based on a subset and the data in the table is highly aggregated data and should be interpreted with caution.

CMP: There are substantially more websites using the CMP Cookiebot in our sample than websites using OneTrust. We control for the CMP as their interface might influence how easy it is for a website to change its cookie settings to become compliant. The language and advertisement of a CMP might also influence the compliance perception of a website. They might have to different degrees a narrative suggesting that using the CMP already prevents potential non-compliance which then influences how our notification is perceived.

Baseline: Websites with few compliance issues to begin with might react differently to the treatment than websites with many compliance issues. Thus, we include a baseline number of cookies as an endogenous covariate.[2] The start rate is calculated as the total number of unique flagged cookies before the treatment.

Response Rate: Of the remaining sample 21 websites have contacted us. This is a response rate of 0.9% after the first two weeks of the experiment. We expect this to slightly increase after the notification message.

6.1.2 Reliability of the Crawler

A main challenge to conducting comprehensive compliance studies is fluctuations over the crawls. This section aims to disentangle the random effects and side effects in the crawls to better understand our data. Overall we analyze 15 time points before the treatment and 13 after the treatment. This

issues on all their websites (i.e. recieve one email per website) while ensuring they receive only one type of treatment. That is why we assigned same email addresses the same treatment after an initial random treatment assignment.

[2] A predictor variable that is influenced or determined by other variables within the same model.

	Control	Self-Check	Report	Self-C.&Rep.	Sum
Germany	307	291	308	293	1199
France	89	84	84	90	347
Ireland	199	197	189	199	784
Sum	595	572	581	582	2330

Table 6.1: Number of websites per country per treatment.

	OneTrust	Cookiebot	Sum
Germany	115	1084	1199
France	43	304	347
Ireland	155	629	784
sum	313	2017	2330

Table 6.2: Number of websites per country per CMP.

Country	Average Total Cookies
Germany	10.12
France	12.80
Ireland	13.25

Table 6.3: Average number of cookies per URL per country before treatment.

represents a preliminary analysis of the findings as the experiment is still running by the time this thesis is written. The output data shows that the crawler has a high variance regarding how many websites are crawled daily and how many cookies are found per website.

Variance in Number of Websites: First, we want to analyze for how many websites the crawler consistently produces. Fig. 6.1 shows that the variance is high. For example, on the 1st of July the crawler only gathered cookies from 871 websites from our sample while on the 19th of July, it gathered cookies from 2832 websites. It seems that the number of found web pages depends on the day and there is no correlation over subsequent days. Note that after the 24th of July, no crawl gathers more than 2500 websites. This could indicate that after notifying websites our crawler was recognized as such and blocked by several websites. However, - and probably more likely - this pattern can also be random and without a clear cause. As there is more data to come this will be discussed more in further work.

Of our sample of 3003 websites only 678 were crawled every day of the 28 days. Fig. 6.2 shows the frequency distribution of missing crawls per website from our sample. We can see that many websites are irregularly crawled successfully. If the crawling success was fully independent on anything else, it would have an exponential distribution. The long right tail implies that if a website failed once it is more likely to fail again. Potential issues could for

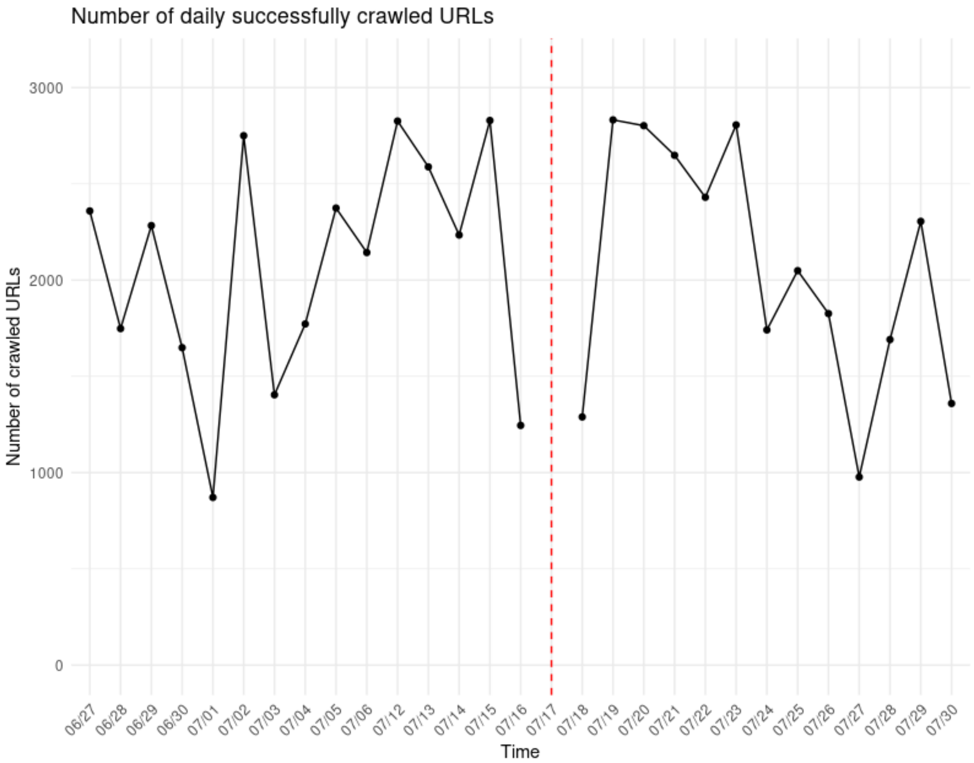

Figure 6.1: The number of successfully crawled websites per day over time. The red dashed line highlights the date of the notification, the 07/17. On this day we omit the crawl data.

example be due to technical glitches or connectivity issues. The prevalence of redirects and frequent changes in URL structures across websites might challenge the crawler's ability to accurately capture content. Additionally, fluctuations in server load and occasional accessibility problems experienced by certain websites could further explain the pattern.

To be able to create a baseline and measure the treatment effect, we decide on a threshold to exclude invalid data. Specifically we subsequently only include websites which have at least one successful crawl per week before the treatment. This allows us to create a baseline of cookie measurements. This threshold reduces the number of observed websites in our sample from 3003 to 2345.

6.1.3 Dependent Variable: Crawl Results

This section describes the data the crawler produces and the baseline before the treatment. The cookies that are recognized by the crawler as the cause of a privacy law violation are referred to as *flagged cookies* (as they were flagged

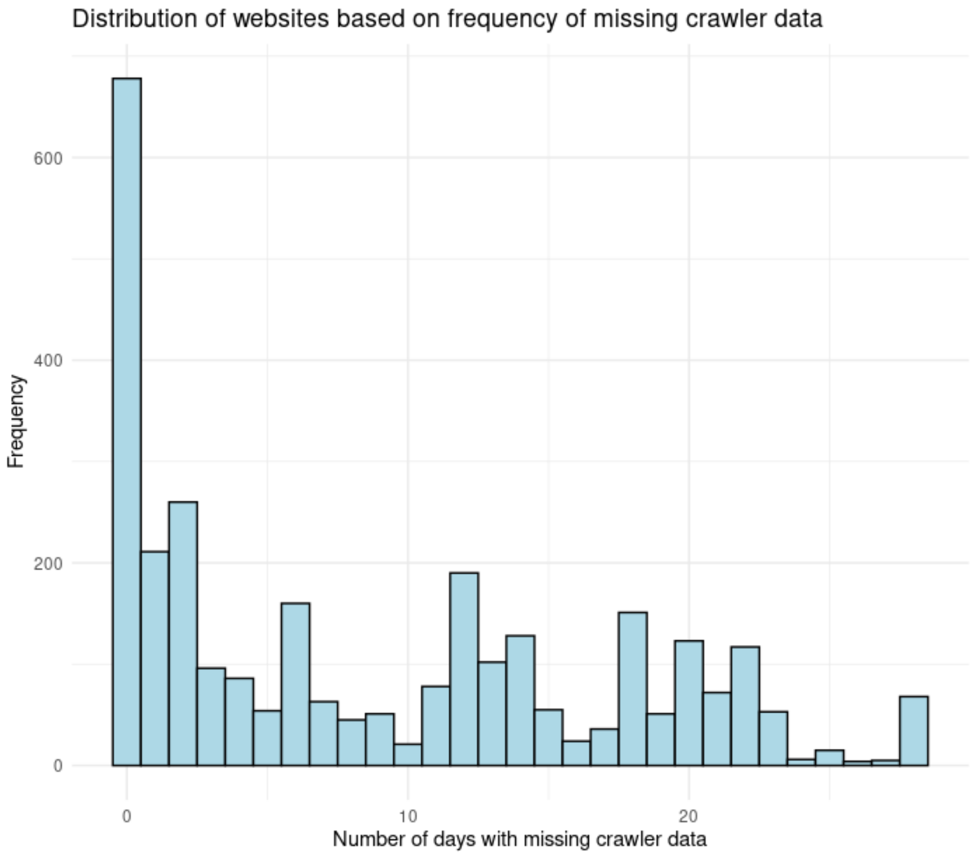

Figure 6.2: The distribution of websites according to how many times the crawler produced no data for them. 678 websites are successfully crawled each time.

by the crawler).

Variance in the Number of Flagged Cookies: In Figure 6.3 we show the (normalized) number of cookies flagged by the crawler over all websites per day for all the days before the treatment. For this plot only websites with at least one successful crawl every week are kept in the dataset. The red line is Just as the number of successfully crawled websites also the number of flagged cookies varies a lot from day to day. Concretely, it varies between 11 867 and 41 510 flagged cookies per day. There also does not seem to occur autocorrelation between the daily crawls. If we compare the number of flagged cookies per day with the number of crawled websites we can see a high correlation (see again Figure 6.3). This is as expected because when more websites are crawled the crawler is likely to find more cookie issues in total. However, the number of successfully crawled websites only partially explains the variance in the total number of cookies.

To better understand the variance in the number of found cookies per website we look at the the the violation categories separately. How many cookies are found on average depends strongly on the type of the cookie violation as Figure 6.4 shows. For the categories `Wrong Expiration Time` and `Wrongly Categorized` in most cases less than five cookies were found per website per crawl and for most websites no cookies at all were found in this category. The category of `Undeclared Cookies` had large differences between websites. Most websites were having no undeclared cookies, while for a few websites more than 100 undeclared cookies were found. Violations in the `Non-Essential Cookies` category were most common. The majority of websites do have at least one found cookie in this category.

Table 6.4 displays the average of the variance per website over time for the control group. Since the control group did not get any treatment, a deterministic crawler would always flag the same number of cookies, i.e. resulting in zero variance. The table shows that `Non-Essential Cookies` violations have the highest variance. Up to some level, this pattern can be expected. High variance can be explained with subpages containing many violations. A certain subpage might be visited on one crawl but not in another resulting in fluctuations on the number of cookies found in a specific website. In comparison, `Wrongly Categorized` has the lowes variance. The most common cookies flagged in this category belong to Google Analytics service, and therefore they are also very often misclassified [4]. This cookie is always registered as soon as the crawler visits the website independent of the specific subpages the crawler visits on the site. This can explain the low variance in this category.

	Variance
Wrong Expiration Time	1.27
Undeclared Cookies	46.96
Wrongly Categorized	0.11
Non-Essential	70.89

Table 6.4: Average of Variance over all dates per cookie issue for the control group.

In regard to the average number of daily flagged cookies over time we expect the number to decrease after the day when the treatment was set. Previous studies suggest that for complex privacy issues websites the remediation rate will be in single digit percentages. Thus, we don't expect a sharp decrease the number of daily flagged cookies. Further, we also expect that for complex privacy issues the time it takes companies to mitigate the issues is is at least a week but likely longer. To clarify, fixing the cookie consent issue itself should not take longer than a day. However, we need to account for communication delay and scheduling overhead within the notified organization. Lastly, we

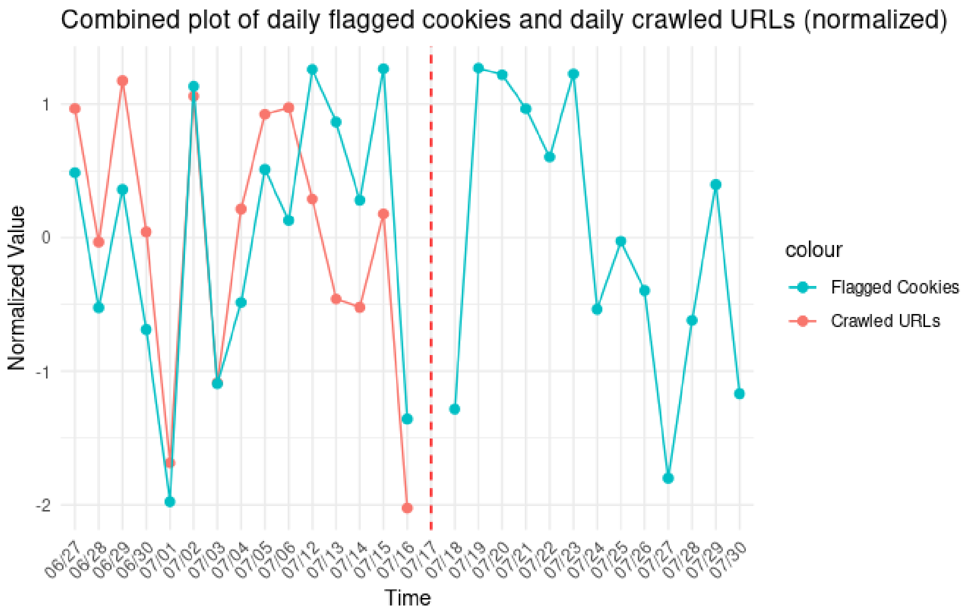

Figure 6.3: Combined Plot of flagged cookies over all websites per day over time pre-treatment and the number of successfully crawled websites per day over time. The red dashed line highlights the date of the notification, the 07/17. On this day we omit the crawl data.

have seen that the crawler varies substantially. Thus, it is not clear if we can actually see a drop in flagged number of cookies, especially on the aggregated level. What is more relevant is a decrease in number of flagged cookies compared to the control group.

The Figures 6.5, 6.6, 6.7, and 6.8 show the average number of cookies for each cookie violation type over time grouped by treatment. As expected immediate decreases in flagged cookies for the treatment groups compared to the control groups aren't evident. The figures illustrate again the high variance of the crawler.

6.2 Responses to the Notifications

We received automated responses that our email was received from 371 websites. Within the first two weeks after the notifications were sent, we received 28 responses regarding the study. Most websites contacted us via the email address dedicated to the study for several different reasons.

Three websites said they were not interested in our tools (audit report or self-check tool). This is interesting as it shows that our website was not perceived as threatening, just like we hoped. Two websites contact us because they needed further assistance to solve problems with their cookies. 3 websites

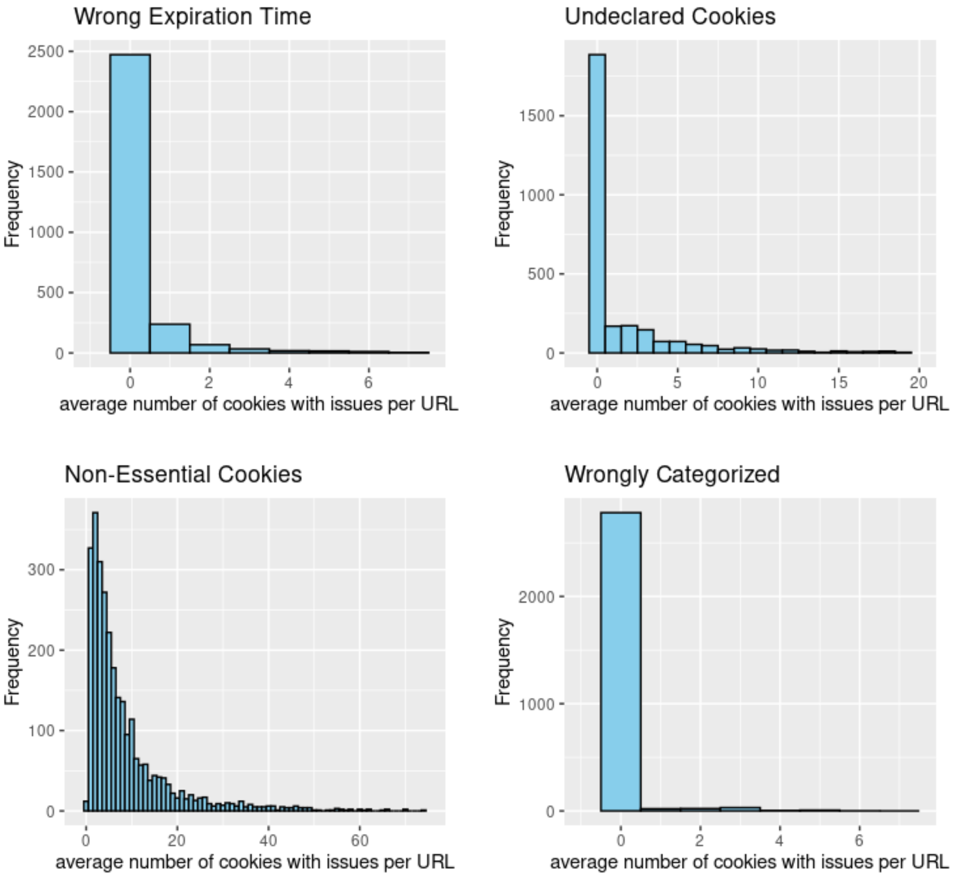

Figure 6.4: The frequency of the found number of cookies for all websites one week before the treatment.

notified us that they were not responsible for the website we contacted them about. A lower bound for the false positive rate is thus 0.01%. Most other emails were notifications that the emails has been received and forwarded internally. Note that those were manual responses.

Two websites wrote specifically to express thanks and made suggestions to improve the audit report. Overall, the tone of the websites was neutral or positive. No negative, aggressive, or threatening response has been received.

6.3 Discussion

This chapter aims to provide an overview of the collected data so far. The data stems from a real-world experiment and as shown, the crawler has strong day-to-day fluctuations. Thus, it is important to understand and disentangle

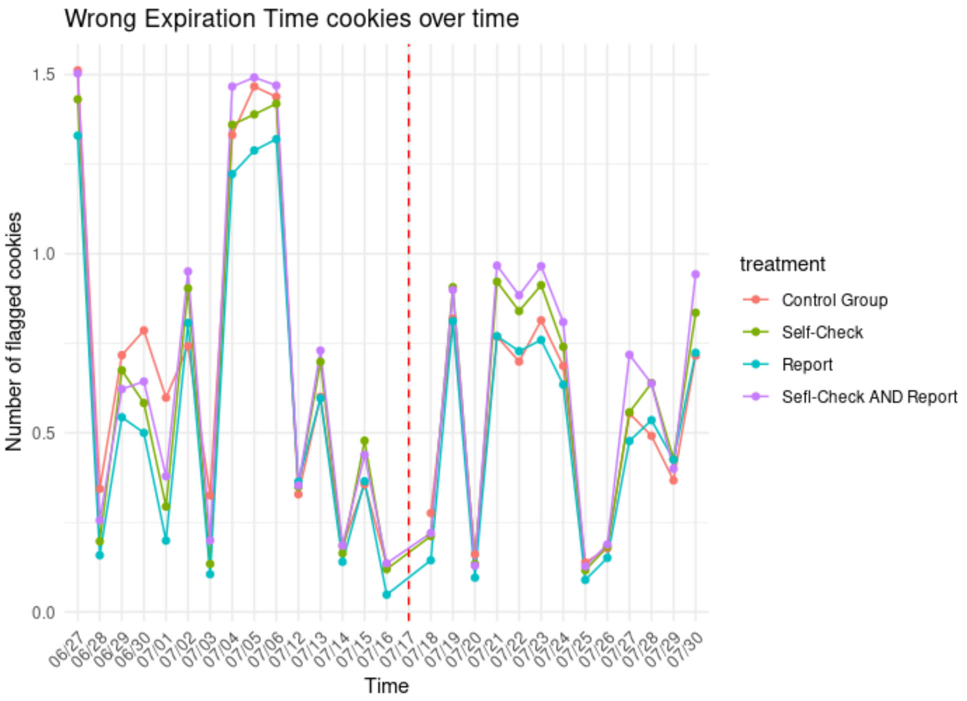

Figure 6.5: The average daily number of flagged cookies in the violation category Wrong Expiration Time. The red dashed line highlights the date of the notification, the 07/17. On this day we omit the crawl data.

the data in detail in a first step. Inferential statistics will be conducted at the end of the experiment.

We have described the data before the treatment and for two weeks after the treatment in this chapter. We found that the reliability of the crawler is low. We outlined possible explanations for the variation in the number of daily visited websites. This includes connectivity issues and fluctuations in the server load. Further, a website might recognize the crawler and block us. To address this issue we suggest including daily fixed effects in the inferential analysis. They can account for the indirect daily effects the crawler has on the output data.

The variance in the number of websites crawled each day is large, with some days yielding far fewer websites than others. The number of daily flagged cookies correlates strongly with the number of visited websites. We argue that the fixed effects can also account for this effect. However, the fluctuation of the number of found cookies can only partially be explained by the crawler's success. If we group the cookies according to their violations different patterns emerge. Non-Essential Cookies violations exhibit the highest

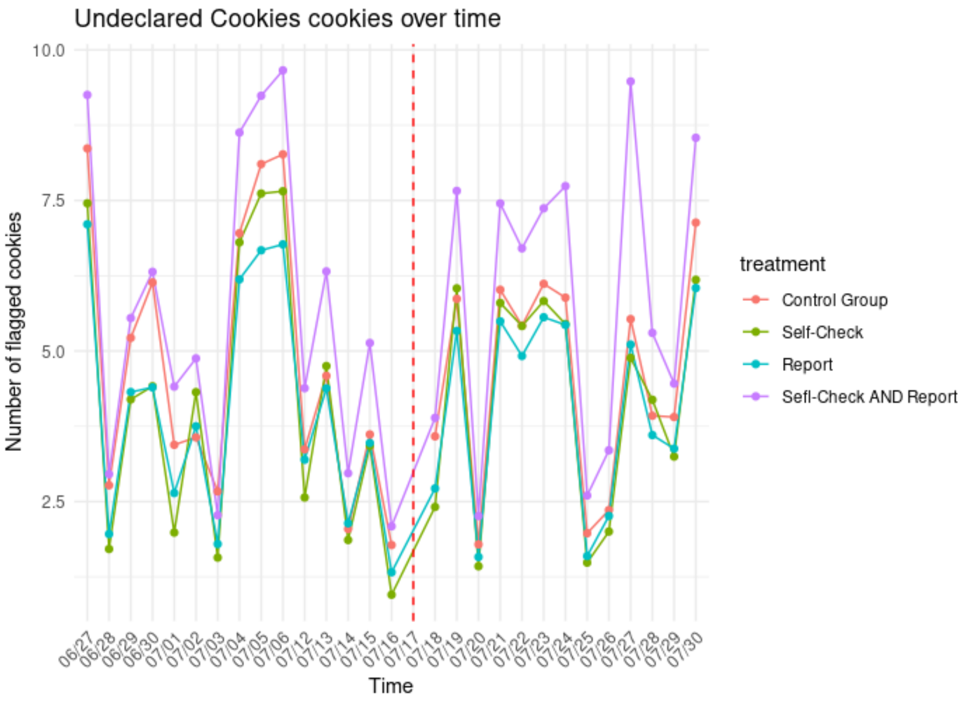

Figure 6.6: The average daily number of flagged cookies in the violation category Undeclared Cookies. The red dashed line highlights the date of the notification, the 07/17. On this day we omit the crawl data.

variance. A possible explanation for the high variance is, that the crawler doesn't always visit the same subpages. Depending on which subpage was visited the number of found cookies can greatly vary.

For the further analysis which will be conducted after the experiments are finished, we suggest several approaches how to deal with the variance for the inferential statistics:

Running Mean: To smooth out the variances a running mean over several days can be used. We suggest using at least five days at a time to account for several weak crawls in a row. This method has been used in notification studies before [21]. However, it does not take into account how many unique cookies the crawler finds.

Union of Cookies: Another approach could be to focus more on the individual cookies that were found instead of aggregating them. We could use a 'running union' similar to a running mean and track the number of unique cookies for intervals after the treatment. Missing runs are treated as if no cookies were found. This approach makes use of more information in our

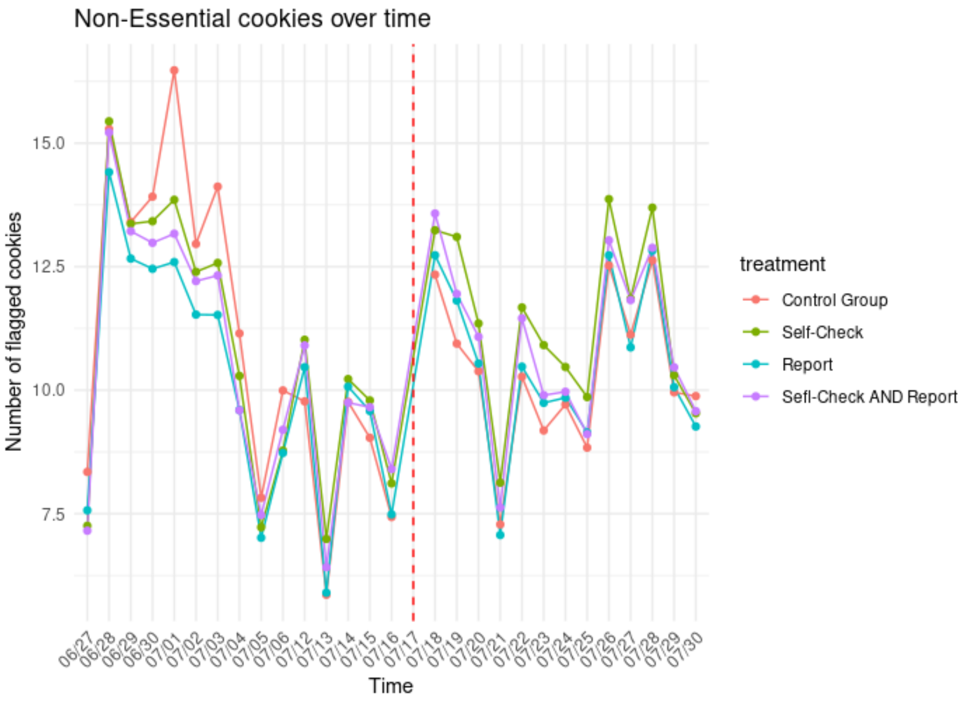

Figure 6.7: The average daily number of flagged cookies in the violation category Non-Essential Cookies. The red dashed line highlights the date of the notification, the 07/17. On this day we omit the crawl data.

data. Further, it can handle days with missing crawls well. However, to implement this approach we need to deal with *dynamic cookies*. Those are cookies that change their name on each visit. There are approaches on how to recognize but might still include manual work.

Baseline Cookies: We only analyze the cookies flagged in the sample run. Those are also the cookies we mention in the audit report. This would be the most rigorous approach to test the effect of the audit report. It is not clear how this measurement captures the effect of the self-audit tool. It also does not account for cases in which websites rename cookies or add new ones. This approach also has to handle dynamic cookies.

Additionally, the experiment contains a reminder email that has not yet been sent out. It also includes a survey, which is not sent out yet. All the responses are coded regarding the sentiment and reason for contact. They further contribute to the understanding of the treatment effects. The experiment contains several sources of data which can provide a conclusive picture together. Despite the high variance in the crawler, we suggest that a meaningful conclusion of the experiment can be made once all data has been

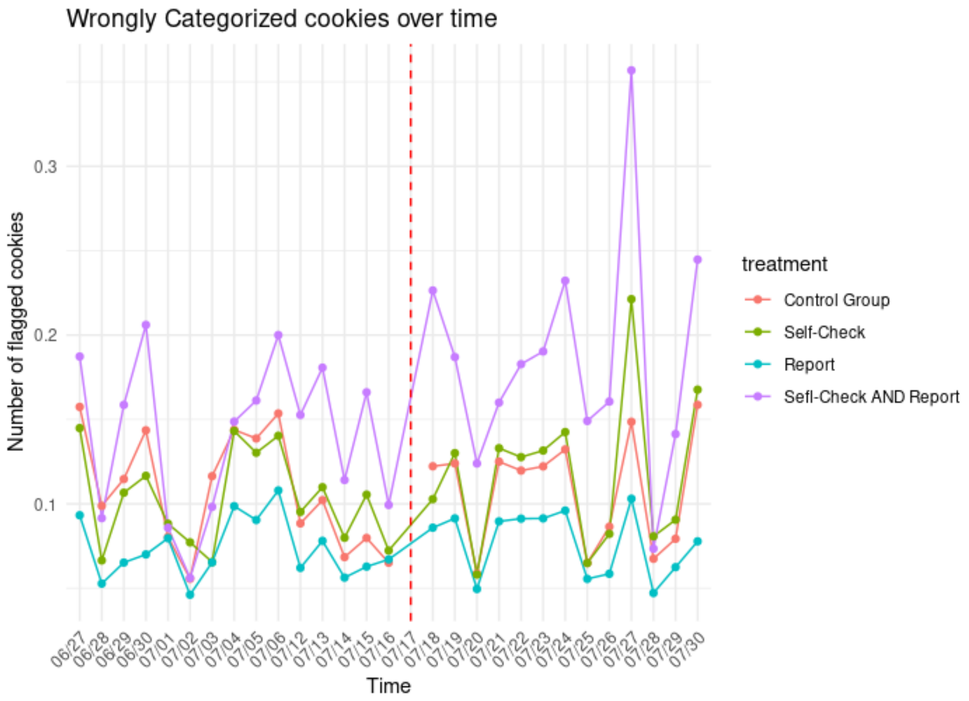

Figure 6.8: The average daily number of flagged cookies in the violation category Wrongly Categorized Cookies. The red dashed line highlights the date of the notification, the 07/17. On this day we omit the crawl data.

gathered.

Chapter 7

Conclusion

The importance of data has grown significantly over the past years. While data used to be a secondary outcome of internet use, it's ownership and proper use have now become central to our lives. As a consequence the EU created several privacy laws, the most comprehensive of which is the General Data Protection Regulation (GDPR). However, a majority of studies find that many websites fail to fully comply with the GDPR [4, 5, 10, 6].

In this context, our goal was to better understand and quantify mechanisms behind cookie consent compliance. We further want to explore the effect of privacy notification including self-empowering tools for websites. We contribute to research on notification theory by including the self-check tool and by notifying about a complex privacy issue with different types of information (legal information, technical information, or both). We suggest three mechanisms behind non-compliance: a gap in legal understanding, a gap in technical understanding, and a gap in translating legal requirements into technical policies. To quantify these mechanisms we create three different types of email notification. The first type contains an audit report to close legal gaps, the second one contains a self-check tool to close technical gaps and the third one contains both to test the interaction effect. In addition, we send out a reminder email and a debriefing email with an invitation to participate and a survey.

We measure the non-compliant behavior of websites with an audit crawler [4]. This crawler is used to create a sample of websites that we contact per email. We run the same crawler daily during the period of the experiment to gather data on potential remediation rates. Note that at the time this thesis is written, the experiment is still running. The reminder and the survey have not yet been sent out. Email responses are encoded for the first two weeks of the experiment. Thus, this thesis does not contain a concrete conclusion of the experiment. Instead, it aims to explain the design of the study in detail and provide first descriptive insights into the crawl data.

We find that the crawler has a high variance in the number of websites it visits daily and in the number of cookies it flags per run. Potential explanations include connectivity issues and fluctuations in the server load. The variance in the number of flagged cookies depends on the type of cookie violation.

So far, we don't see a decrease in the number of flagged cookies for the treatment groups compared to the control groups for all violation types. However, this is not completely unexpected: We estimate low remediation rates as we inform them about complex cookie issues. Further, we only consider data from the first two weeks after the treatment, which might not be enough time for websites to address the issues. The effect of privacy notifications on the GDPR is unclear with the current data. However, the notification email, the survey data, and the crawler data for the full two months are still missing. Thus, no conclusive remarks can be made yet.

This thesis provides insights on how into design a theory-driven notification study. It also highlights and analyzes the challenges in regard to unreliable crawlers. In this context, we propose three specific approaches how to deal with the variance.

Bibliography

[1] European Parliament and Council of the European Union, *Regulation (eu) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*, Last accessed on: 27.06.2023, 2016. [Online]. Available: `http://data.europa.eu/eli/reg/2016/679/2016-05-04`.

[2] Council of the European Union and European Parliament. "Directive 2002/58/ec of the european parliament and of the council of 12 july 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (directive on privacy and electronic communications)." (2002), [Online]. Available: `https://eur-lex.europa.eu/eli/dir/2002/58/oj`.

[3] D. Bollinger, "Analyzing cookies compliance with the gdpr," M.S. thesis, ETH Zurich, Department of Computer Science, 2021. [Online]. Available: `https://doi.org/10.3929/ethz-b-000477333`.

[4] D. Bollinger, K. Kubicek, C. Cotrini, and D. Basin, "Automating cookie consent and GDPR violation detection," in *31st USENIX Security Symposium (USENIX Security 22)*, Boston, MA: USENIX Association, Aug. 2022, pp. 2893–2910, ISBN: 978-1-939133-31-1. [Online]. Available: `https://www.usenix.org/conference/usenixsecurity22/presentation/bollinger`.

[5] A. Bouhoula, "Automated detection of gdpr violations in cookie notices using machine learning," M.S. thesis, ETH Zurich, Department of Computer Science, 2022. [Online]. Available: `https://doi.org/10.3929/ethz-b-000575741`.

[6] C. Matte, N. Bielova, and C. Santos, *Do cookie banners respect my choice? measuring legal compliance of banners from iab europe's transparency and consent framework*, IEEE, 2020. arXiv: `1911.09964 [cs.CR]`.

[7] S. Sirur, J. R. Nurse, and H. Webb, "Are we there yet? understanding the challenges faced in complying with the general data protection regulation (gdpr)," in *Proceedings of the 2nd International Workshop on Multimedia Privacy and Security*, ser. MPS '18, Toronto, Canada: Association for Computing Machinery, 2018, pp. 88–95, ISBN: 9781450359887. DOI: 10.1145/3267357.3267368. [Online]. Available: https://doi.org/10.1145/3267357.3267368.

[8] M. Freitas and M. Mira da Silva, "Gdpr compliance in smes: There is much to be done," *Journal of Information Systems Engineering & Management*, vol. 3, Nov. 2018. DOI: 10.20897/jisem/3941.

[9] Z. S. Li, C. Werner, N. Ernst, and D. Damian, "Towards privacy compliance: A design science study in a small organization," *Information and Software Technology*, vol. 146, p. 106 868, 2022, ISSN: 0950-5849. DOI: https://doi.org/10.1016/j.infsof.2022.106868. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950584922000362.

[10] C. Utz, M. Degeling, S. Fahl, F. Schaub, and T. Holz, "(un)informed consent: Studying gdpr consent notices in the field," *CoRR*, vol. abs/1909.02638, 2019.

[11] T. Sakamoto and M. Matsunaga, "After gdpr, still tracking or not? understanding opt-out states for online behavioral advertising," *2019 IEEE Security and Privacy Workshops (SPW)*, pp. 92–99, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:202701186.

[12] I. Sánchez-Rola *et al.*, "Can i opt out yet?: Gdpr and the global illusion of cookie control," *Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security*, pp. 340–351, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:195848775.

[13] X. Hu and N. R. Sastry, "Characterising third party cookie usage in the eu after gdpr," *Proceedings of the 10th ACM Conference on Web Science*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:145050938.

[14] M. Nouwens, I. Liccardi, M. Veale, D. Karger, and L. Kagal, "Dark patterns after the GDPR: Scraping consent popups and demonstrating their influence," *CoRR*, vol. abs/2001.02479, 2020.

[15] T. T. Nguyen, M. Backes, and B. Stock, "Freely given consent? studying consent notice of third-party tracking and its violations of gdpr in android apps," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '22, Los Angeles, CA, USA: Association for Computing Machinery, 2022, pp. 2369–2383, ISBN: 9781450394505. DOI: 10.1145/3548606.3560564. [Online]. Available: https://doi.org/10.1145/3548606.3560564.

[16] V. Ayala-Rivera and L. Pasquale, "The grace period has ended: An approach to operationalize gdpr requirements," in *2018 IEEE 26th International Requirements Engineering Conference (RE)*, 2018, pp. 136–146. DOI: 10.1109/RE.2018.00023.

[17] J. Mohan, M. Wasserman, and V. Chidambaram, "Analyzing GDPR compliance through the lens of privacy policy," *CoRR*, vol. abs/1906.12038, 2019. arXiv: 1906.12038. [Online]. Available: http://arxiv.org/abs/1906.12038.

[18] F. Schaub, R. Balebako, A. L. Durity, and L. F. Cranor, "A design space for effective privacy notices," in *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, Ottawa: USENIX Association, Jul. 2015, pp. 1–17, ISBN: 978-1-931971-249. [Online]. Available: https://www.usenix.org/conference/soups2015/proceedings/presentation/schaub.

[19] J. Gluck *et al.*, "How short is too short? implications of length and framing on the effectiveness of privacy notices," in *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, Denver, CO: USENIX Association, Jun. 2016, pp. 321–340, ISBN: 978-1-931971-31-7. [Online]. Available: https://www.usenix.org/conference/soups2016/technical-sessions/presentation/gluck.

[20] M. Maass *et al.*, "Effective notification campaigns on the web: A matter of trust, framing, and support," in *30th USENIX Security Symposium (USENIX Security 21)*, USENIX Association, Aug. 2021, pp. 2489–2506, ISBN: 978-1-939133-24-3. [Online]. Available: https://www.usenix.org/conference/usenixsecurity21/presentation/maass.

[21] C. Utz, M. Michels, M. Degeling, N. Marnau, and B. Stock, "Comparing large-scale privacy and security notifications," in *PETS 2023*, Jul. 2023. [Online]. Available: https://publications.cispa.saarland/3918/.

[22] A. Hennig, H. Dietmann, F. Lehr, M. Mutter, M. Volkamer, and P. Mayer, ""your cookie disclaimer is not in line with the ideas of the gdpr. why?"" In *Human Aspects of Information Security and Assurance*, N. Clarke and S. Furnell, Eds., Cham: Springer International Publishing, 2022, pp. 218–227, ISBN: 978-3-031-12172-2.

[23] M. Hils, D. W. Woods, and R. Böhme, "Measuring the emergence of consent management on the web," in *Proceedings of the ACM Internet Measurement Conference*, ser. IMC '20, Virtual Event, USA: Association for Computing Machinery, 2020, pp. 317–332, ISBN: 9781450381383. DOI: 10.1145/3419394.3423647. [Online]. Available: https://doi.org/10.1145/3419394.3423647.

[24] F. Lancieri, "Narrowing data protection's enforcement gap," *Maine Law Review*, vol. 74, no. 1, 2022, Available at SSRN: https://ssrn.com/abstract=3806880 or http://dx.doi.org/10.2139/ssrn.3806880.

[25] G. S. Becker, "Crime and punishment: An economic approach," *Journal of Political Economy*, vol. 76, no. 2, p. 169, 1968.

[26] K. Kubicek, D. Bollinger, A. Zanga, C. Cotrini, and D. Basin, *CookieBlock & CookieAudit: Fixing cookie consent with ML*, https://github.com/Fredilein/CookieAudit, Boston, MA, Aug. 2022.

[27] K. Ruth, D. Kumar, B. Wang, L. Valenta, and Z. Durumeric, "Toppling top lists: Evaluating the accuracy of popular website lists," in *Proceedings of the 22nd ACM Internet Measurement Conference*, ser. IMC '22, Nice, France: Association for Computing Machinery, 2022, pp. 374–387, ISBN: 9781450392594. DOI: 10.1145/3517745.3561444. [Online]. Available: https://doi.org/10.1145/3517745.3561444.

[28] Alexa. "We will be retiring the alexa.com apis on december 15, 2022." (2022), [Online]. Available: https://support.alexa.com/hc/en-us/articles/4411466276375.

[29] B. Krumnow, H. Jonker, and S. Karsch, "How gullible are web measurement tools? a case study analysing and strengthening openwpm's reliability," in *Proceedings of the 18th International Conference on Emerging Networking EXperiments and Technologies*, ser. CoNEXT '22, Roma, Italy: Association for Computing Machinery, 2022, pp. 171–186, ISBN: 9781450395083. DOI: 10.1145/3555050.3569131. [Online]. Available: https://doi.org/10.1145/3555050.3569131.

[30] Article 29 Working Party, European Union, *Guidelines on Consent under Regulation 2016/679 (WP259 rev.01)*, https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=623051; Last accessed on: 2021.02.06, 2018.

Appendix A

Email Texts

A.1 First Contact Email

A.1.1 Treatment Self-Check

English Version:

Dear Madam or Sir,

We are a legal-computer science research group from the Swiss Federal Institute of Technology in Zurich (ETH Zurich) and are contacting you because you are listed in the imprint of the following website as the responsible party: url

We are conducting a scientific study to better understand how websites comply with privacy laws according to the European General Data Protection Regulation (GDPR) and the EU ePrivacy Directive.

As part of a research project on internet privacy, we have developed a browser extension that can help website owners improve their compliance with privacy laws called CookieAudit [link]. The extension allows you to identify potential privacy issues and informs you how to address these. It detects consent, used cookies, and reports potential privacy issues. CookieAudit is provided and operated by the Information Security Group at ETH Zurich [link]. It is free to use and there is no obligation to make any changes to your website based on the results.

If you have any questions, please feel free to contact us via [Study Email Address]. You can find more information about our research project on the ETH website. We thank you for your attention to this matter.

Sincerely, Laura-Vanessa Soldner (Student ETH Zurich),
Karel Kubicek (ETH Zurich Information Security Group),
Amit Zac (ETH Zurich Center for Law & Economics)

German Version:

Sehr geehrte Damen und Herren

Wir sind eine interdisziplinäre Forschungsgruppe mit Fokus auf Informatik und Recht der Eidgenössischen Technischen Hochschule Zürich (ETH) und kontaktieren Sie, da Sie im Impressum der folgenden Webseite als Verantwortliche/r aufgeführt sind: url.

Der Schwerpunkt unserer Forschung liegt darin, besser zu verstehen, ob und wieso Webseiten die Datenschutzgesetze gemäß der Europäischen Datenschutz-Grundverordnung (GDPR) und der EU-Datenschutzrichtlinie für elektronische Kommunikation einhalten.

Im Rahmen eines Forschungsprojekts zum Datenschutz im Internet haben wir eine Browser-Erweiterung namens CookieAudit entwickelt, die Webseiten-Betreibern helfen kann, die Einhaltung der Datenschutzgesetze zu gewährleisten [Link]. Die Browser-Erweiterung ermöglicht es Ihnen, potenzielle Verstösse zu erkennen und informiert Sie darüber, wie Sie diese beheben können. Das Tool erkennt die Zustimmung zu den Cookies, die verwendeten Cookies und erstellt einen Bericht über mögliche Nichteinhaltung der Datenschutzgesetze. CookieAudit wird von der Forschungsgruppe Information Security der ETH Zürich bereitgestellt und betrieben [Link]. Die Nutzung von CookieAudit ist kostenlos und es besteht keine Verpflichtung, aufgrund der Ergebnisse Änderungen an Ihrer Website vorzunehmen.

Falls Sie Fragen haben, können Sie uns gerne über [Study Email Address] kontaktieren. Mehr Informationen über unser Forschungsprojekt finden Sie auf der Webseite der ETH. Wir danken Ihnen für Ihre Aufmerksamkeit in dieser Angelegenheit.

Mit freundlichen Grüssen

Laura-Vanessa Soldner (Studentin ETH Zürich),
Karel Kubicek (ETH Zürich Information Security Group),
Amit Zac (ETH Zürich Center for Law & Economics)

French Version:

Madame, Monsieur,

Nous sommes un groupe de recherche en informatique juridique de l'École polytechnique fédérale de Zurich (ETH Zurich) et nous vous contactons parce que vous figurez dans les mentions légales du site web suivant en tant que partie responsable : url

Nous menons une étude scientifique pour mieux comprendre comment les sites web se conforment aux lois sur la protection de la vie privée conformément au règlement général européen sur la protection des données

(RGPD) et à la directive européenne sur la vie privée et les communications électroniques.

Dans le cadre d'un projet de recherche sur la protection de la vie privée sur l'internet, nous avons développé une extension de navigateur qui peut aider les propriétaires de sites web à mieux respecter les lois sur la protection de la vie privée, appelée CookieAudit [lien]. Cette extension vous permet d'identifier les problèmes potentiels en matière de protection de la vie privée et vous informe sur la manière de les résoudre. Elle détecte le consentement, les cookies utilisés et signale les problèmes potentiels en matière de protection de la vie privée. CookieAudit est fourni et géré par le groupe de sécurité de l'information de l'ETH Zurich [lien]. Son utilisation est gratuite et il n'y a aucune obligation d'apporter des modifications à votre site web sur la base des résultats.

Si vous avez des questions, n'hésitez pas à nous contacter via [Study Email Address]. Vous trouverez plus d'informations sur notre projet de recherche sur le site de l'ETH. Nous vous remercions de votre temps.

Nous vous prions d'agréer, Madame, Monsieur, l'expression de nos salutations distinguées,

Laura-Vanessa Soldner (étudiante à l'ETH Zurich),
Karel Kubicek (ETH Zurich Information Security Group),
Amit Zac (ETH Zurich Center for Law & Economics)

A.1.2 Treatment Report

English Version:

Dear Madam or Sir,

We are a legal-computer science research group from the Swiss Federal Institute of Technology in Zurich (ETH Zurich) and are contacting you because you are listed in the imprint of the following website as the responsible party: url

We are conducting a scientific study to better understand how websites comply with privacy laws according to the European General Data Protection Regulation (GDPR) and the EU ePrivacy Directive.

As part of a research project on internet privacy, we are analysing various websites on how they implement privacy laws. Our analysis has revealed that there may be issues in how your website addresses privacy laws. Specifically, we have observed that your website might not handle cookie notices and personal data collection correctly. Please find a detailed report of our findings attached as a PDF. We provide our preliminary findings to you as a courtesy and, as a research team at a university, will not reveal such individual findings to anyone. We are a research institute and cannot offer any advice on the legality of your actions.

If you have any questions, please feel free to contact us via [Study Email Address]. You can find more information about our research project on the ETH website. We thank you for your attention to this matter.

Sincerely, Laura-Vanessa Soldner (Student ETH Zurich),
Karel Kubicek (ETH Zurich Information Security Group),
Amit Zac (ETH Zurich Center for Law & Economics)

German Version:

Sehr geehrte Damen und Herren

Wir sind eine interdisziplinäre Forschungsgruppe mit Fokus auf Informatik und Recht der Eidgenössischen Technischen Hochschule Zürich (ETH) und kontaktieren Sie, da Sie im Impressum der folgenden Webseite als Verantwortliche/r aufgeführt sind: url.

Der Schwerpunkt unserer Forschung liegt darin, besser zu verstehen, ob und wieso Webseiten die Datenschutzgesetze gemäß der Europäischen Datenschutz-Grundverordnung (GDPR) und der EU-Datenschutzrichtlinie für elektronische Kommunikation einhalten.

Im Rahmen eines Forschungsprojekts zum Datenschutz im Internet analysieren wir verschiedene Webseiten daraufhin, wie sie das Datenschutzrecht umsetzen. Unsere Analyse hat ergeben, dass Ihre Webseite möglicherweise nicht alle Regelungen der Datenschutzgesetze einhält. Insbesondere haben wir festgestellt, dass Ihre Webseite potentiell nicht korrekt mit Cookie-Hinweisen und der Erhebung personenbezogener Daten umgeht. Einen ausführlichen Bericht über unsere Feststellungen finden Sie im Anhang als PDF. Wir stellen Ihnen unsere vorläufigen Ergebnisse aus Höflichkeit zur Verfügung und werden als universitäres Forschungsteam individuelle Ergebnisse niemandem gegenüber offenlegen. Weiter möchten wir Sie darauf hinweisen, dass wir ein Forschungsinstitut sind und keine Rechtsberatung anbieten können.

Falls Sie Fragen haben, können Sie uns gerne über [Study Email Address]. kontaktieren. Mehr Informationen über unser Forschungsprojekt finden Sie auf der Webseite der ETH. Wir danken Ihnen für Ihre Aufmerksamkeit in dieser Angelegenheit.

Mit freundlichen Grüssen

Laura-Vanessa Soldner (Studentin ETH Zürich),
Karel Kubicek (ETH Zürich Information Security Group),
Amit Zac (ETH Zürich Center for Law & Economics)

French Version:

Madame, Monsieur,

Nous sommes un groupe de recherche en informatique juridique de l'École polytechnique fédérale de Zurich (ETH Zurich) et nous vous contactons parce que vous figurez dans les mentions légales du site web suivant en tant que partie responsable : url

Nous menons une étude scientifique pour mieux comprendre comment les sites web se conforment aux lois sur la protection de la vie privée conformément au règlement général européen sur la protection des données (RGPD) et à la directive européenne sur la vie privée et les communications électroniques.

Dans le cadre d'un projet de recherche sur la protection de la vie privée sur l'internet, nous analysons différents sites web sur la manière dont ils appliquent les lois sur la protection de la vie privée. Notre analyse a révélé qu'il pourrait y avoir des problèmes dans la manière dont votre site web traite les lois sur la protection de la vie privée. Plus précisément, nous avons observé que votre site web pourrait ne pas gérer correctement les avis de cookies et la collecte de données personnelles. Vous trouverez un rapport détaillé de nos conclusions en pièce jointe au format PDF. Nous vous communiquons nos conclusions préliminaires par courtoisie et, en tant qu'équipe de recherche universitaire, nous ne divulguerons ces conclusions individuelles à personne. Nous sommes un institut de recherche et ne pouvons offrir aucun conseil sur la légalité de vos actions.

Si vous avez des questions, n'hésitez pas à nous contacter via [Study Email Address]. Vous trouverez plus d'informations sur notre projet de recherche sur le site de l'ETH. Nous vous remercions de votre temps.

Nous vous prions d'agréer, Madame, Monsieur, l'expression de nos salutations distinguées,

Laura-Vanessa Soldner (étudiante à l'ETH Zurich),
Karel Kubicek (ETH Zurich Information Security Group),
Amit Zac (ETH Zurich Center for Law & Economics)

A.1.3 Treatment Self-Check and Report

English Version:

Dear Madam or Sir,

We are a legal-computer science research group from the Swiss Federal Institute of Technology in Zurich (ETH Zurich) and are contacting you because you are listed in the imprint of the following website as the responsible party: url

We are conducting a scientific study to better understand how websites comply with privacy laws according to the European General Data Protection Regulation (GDPR) and the EU ePrivacy Directive.

As part of a research project on internet privacy, we are analysing various websites on how they implement privacy laws. Our analysis has revealed that there may be issues in how your website addresses privacy laws. Specifically, we have observed that your website might not handle cookie notices and personal data collection correctly. Please find a detailed report of our findings attached as a PDF. We provide our preliminary findings to you as a courtesy and, as a research team at a university, will not reveal such individual findings to anyone. We are a research institute and cannot offer any advice on the legality of your actions.

Further, we have developed a browser extension that can help website owners improve their compliance with privacy laws called CookieAudit [link]. The extension allows you to identify potential privacy issues and informs you how to address these. It detects consent, used cookies, and reports potential privacy issues. CookieAudit is provided and operated by the Information Security Group at ETH Zurich [link]. It is free to use and there is no obligation to make any changes to your website based on the results. The report attached to this email has been generated with CookieAudit. We would like to point out two things: Since CookieAudit randomly scans some subpages, the mentioned cookie names may not correspond exactly to those of the report attached to this email. Furthermore, in order to find all cookies in the "Non-essential cookies" category, it is necessary to select the option to reject all cookies in the cookie consent pop-up.

If you have any questions, please feel free to contact us via [Study Email Address]. You can find more information about our research project on the ETH website. We thank you for your attention to this matter.

Sincerely, Laura-Vanessa Soldner (Student ETH Zurich),
Karel Kubicek (ETH Zurich Information Security Group),
Amit Zac (ETH Zurich Center for Law & Economics)

German Version:

Sehr geehrte Damen und Herren

Wir sind eine interdisziplinäre Forschungsgruppe mit Fokus auf Informatik und Recht der Eidgenössischen Technischen Hochschule Zürich (ETH) und kontaktieren Sie, da Sie im Impressum der folgenden Webseite als Verantwortliche/r aufgeführt sind: url.

Der Schwerpunkt unserer Forschung liegt darin, besser zu verstehen, ob und wieso Webseiten die Datenschutzgesetze gemäß der Europäischen

Datenschutz-Grundverordnung (GDPR) und der EU-Datenschutzrichtlinie für elektronische Kommunikation einhalten.

Im Rahmen eines Forschungsprojekts zum Datenschutz im Internet analysieren wir verschiedene Webseiten daraufhin, wie sie das Datenschutzrecht umsetzen. Unsere Analyse hat ergeben, dass Ihre Webseite möglicherweise nicht alle Regelungen der Datenschutzgesetze einhält. Insbesondere haben wir festgestellt, dass Ihre Webseite potentiell nicht korrekt mit Cookie-Hinweisen und der Erhebung personenbezogener Daten umgeht. Einen ausführlichen Bericht über unsere Feststellungen finden Sie im Anhang als PDF. Wir stellen Ihnen unsere vorläufigen Ergebnisse aus Höflichkeit zur Verfügung und werden als universitäres Forschungsteam individuelle Ergebnisse niemandem gegenüber offenlegen. Weiter möchten wir Sie darauf hinweisen, dass wir ein Forschungsinstitut sind und keine Rechtsberatung anbieten können.

Des Weiteren haben wir eine Browser-Erweiterung namens CookieAudit entwickelt, die Webseiten-Betreibern helfen kann, die Einhaltung der Datenschutzgesetze zu gewährleisten [Link]. Die Browser-Erweiterung ermöglicht es Ihnen, potenzielle Verstösse zu erkennen und informiert Sie darüber, wie Sie diese beheben können. Das Tool erkennt die Zustimmung zu den Cookies, die verwendeten Cookies und erstellt einen Bericht über mögliche Nichteinhaltung der Datenschutzgesetze. CookieAudit wird von der Forschungsgruppe Information Security der ETH Zürich bereitgestellt und betrieben [Link]. Die Nutzung von CookieAudit ist kostenlos und es besteht keine Verpflichtung, aufgrund der Ergebnisse Änderungen an Ihrer Website vorzunehmen. Der Bericht am Ende dieses E-Mails wurde mit CookieAudit erstellt. Wir möchten auf zwei Punkte hinweisen: Da CookieAudit einiger Unterseiten zufällig durchsucht, können die genannten Cookie-Namen möglicherweise nicht mit dem Bericht im Anhang übereinstimmen. Um alle Cookies in der Kategorie "Non-essential cookies" zu finden, ist es ausserdem notwendig, im Cookie-Consent-Popup die Option auszuwählen, alle Cookies abzulehnen.

Falls Sie Fragen haben, können Sie uns gerne über [Study Email Address] kontaktieren. Mehr Informationen über unser Forschungsprojekt finden Sie auf der Webseite der ETH. Wir danken Ihnen für Ihre Aufmerksamkeit in dieser Angelegenheit.

Mit freundlichen Grüssen

Laura-Vanessa Soldner (Studentin ETH Zürich),
Karel Kubicek (ETH Zürich Information Security Group),
Amit Zac (ETH Zürich Center for Law & Economics)

French Version:

Madame, Monsieur,

Nous sommes un groupe de recherche en informatique juridique de l'École polytechnique fédérale de Zurich (ETH Zurich) et nous vous contactons parce que vous figurez dans les mentions légales du site web suivant en tant que partie responsable : url

Nous menons une étude scientifique pour mieux comprendre comment les sites web se conforment aux lois sur la protection de la vie privée conformément au règlement général européen sur la protection des données (RGPD) et à la directive européenne sur la vie privée et les communications électroniques.

Dans le cadre d'un projet de recherche sur la protection de la vie privée sur Internet, nous analysons différents sites web sur la manière dont ils appliquent les lois sur la protection de la vie privée. Notre analyse a révélé que la manière dont votre site web applique les lois sur la protection de la vie privée peut poser des problèmes. Plus précisément, nous avons observé que votre site web pourrait ne pas gérer correctement les notifications de cookies et la collecte de données personnelles. Vous trouverez un rapport détaillé de nos conclusions en pièce jointe au format PDF. Nous vous communiquons nos conclusions préliminaires par courtoisie et, en tant qu'équipe de recherche universitaire, nous ne divulguerons ces conclusions individuelles à personne. Nous sommes un institut de recherche et ne pouvons offrir aucun conseil sur la légalité de vos actions.

Par ailleurs, nous avons développé une extension de navigateur qui peut aider les propriétaires de sites web à améliorer leur conformité avec les lois sur la protection de la vie privée, appelée CookieAudit [lien]. Cette extension vous permet d'identifier les problèmes potentiels en matière de protection de la vie privée et vous informe sur la manière de les résoudre. Elle détecte le consentement, les cookies utilisés et signale les problèmes potentiels de protection de la vie privée. CookieAudit est fourni et géré par le groupe de sécurité de l'information de l'ETH Zurich [lien]. Son utilisation est gratuite et il n'y a aucune obligation d'apporter des modifications à votre site web sur la base des résultats. Le rapport joint à cet e-mail a été généré avec CookieAudit. Nous aimerions souligner deux choses : CookieAudit analysant de manière aléatoire certaines sous-pages, les noms de cookies mentionnés peuvent ne pas correspondre exactement à ceux du rapport joint à cet e-mail. En outre, pour trouver tous les cookies dans la catégorie "Non-essential cookies", il est nécessaire de sélectionner l'option de rejeter tous les cookies dans la fenêtre contextuelle de consentement aux cookies.

Si vous avez des questions, n'hésitez pas à nous contacter via [Study Email Address]. Vous trouverez plus d'informations sur notre projet de recherche sur le site de l'ETH. Nous vous remercions de votre temps.

Nous vous prions d'agréer, Madame, Monsieur, l'expression de nos salutations distinguées,

Laura-Vanessa Soldner (étudiante à l'ETH Zurich),
Karel Kubicek (ETH Zurich Information Security Group),
Amit Zac (ETH Zurich Center for Law & Economics)

A.2 Reminder Email

English Version: Dear Madam or Sir,

We contacted you on [DATE] regarding research conducted at the Swiss Federal Institute of Technology in Zurich (ETH Zurich) about how websites comply with privacy laws according to the European General Data Protection Regulation (GDPR) and the EU ePrivacy Directive.

[Treatment A, self-check] We would like to follow up on our email from the [DATE] with a reminder about the browser extension we have developed called CookieAudit [link]. The extension allows you to identify potential privacy issues and informs you how to address these. It detects consent, used cookies, and reports potential privacy issues. CookieAudit is provided and operated by the Information Security Group at ETH Zurich [link]. It is free to use and there is no obligation to make any changes to your website based on the results.

If you have already used this extension, please feel free to ignore this email.

[Treatment B, notice] We would like to follow up on our email from the [DATE] with a reminder on the report that we provided you in this email. This analysis has revealed that there may be issues in how your website addresses privacy laws. Specifically, we have observed that your website might not handle cookie notices and personal data collection correctly. Note that we will not reveal such individual findings to anyone. We are a research institute and cannot offer any advice on the legality of your actions.

We have attached the same report to this email again. Please note that this report displays the same information as the last one. The report was not updated in the meantime. If you have already taken steps to address the potential issues raised in the report we apologise to contact you again, please feel free to ignore this email.

[Treatment C, notice and self-check] We would like to follow up on our email from the [DATE] with a reminder on the PDF report that we provided you in this email. This analysis has revealed that there may be issues in how your website addresses privacy laws. Specifically, we have observed that your website might not handle cookie notices and personal data collection correctly. Note that we will not reveal such individual findings to anyone. We are a research institute and cannot offer any advice on the legality of your actions. We have attached the same report to this email as a PDF again. Please note that this report displays the same information as the last one.

The report was not updated in the meantime. If you have already taken steps to address the potential issues raised in the report we apologise to contact you again, please feel free to ignore this email.

Furthermore, we would like to point out again the browser extension we have developed called CookieAudit [link]. The extension allows you to identify potential privacy issues and informs you how to address these. It detects consent, used cookies, and reports potential privacy issues. CookieAudit is provided and operated by the Information Security Group at ETH Zurich [link]. It is free to use and there is no obligation to make any changes to your website based on the results.The report attached to this email has been generated with CookieAudit. We would like to point out two things: Since CookieAudit randomly scans some subpages, the mentioned cookie names may not correspond exactly to those of the report attached to this email. Furthermore, in order to find all cookies in the ”Non-essential cookies” category, it is necessary to select the option to reject all cookies in the cookie consent pop-up.

If you have any questions, please feel free to contact us via [Study Email Address]. You can find more information about our research project on the ETH website.

We thank you for your attention to this matter.

Sincerely,¡¡ Laura-Vanessa Soldner (Student ETH Zurich),
Karel Kubicek (ETH Zurich Information Security Group),
Amit Zac (ETH Zurich Center for Law & Economics)

A.3 Debriefing and Survey Email

A.3.1 Control Group

English Version: Dear Madam or Sir,

We are a legal-computer science research group from the Swiss Federal Institute of Technology in Zurich (ETH Zurich) and are contacting you because you are listed in the imprint of the following website as the responsible party: url

We are conducting a scientific study to better understand how websites comply with privacy laws according to the European General Data Protection Regulation (GDPR) and the EU ePrivacy Directive. As part of this research, we are using automated methods to analyze websites and assess their compliance with cookie consent requirements. Our analysis suggests that your website may not correctly apply the GDPR and the EU ePrivacy Directive. If you would like to learn more please contact us via the email below.

In addition, to gain insights into your website's perspective on the motivations and challenges related to GDPR implementation, we would like to invite your website to participate in our research project by completing the following survey.

Link to survey: [LINK]

Participation is voluntary and completely anonymous. You have the option to withdraw your participation at any time. Completing the survey will take 5 minutes of your time.

If you have any questions, please feel free to contact us via [Study Email Address]]. You can find more information about our research project on the ETH website. We thank you for your attention to this matter.

Sincerely, Laura-Vanessa Soldner (Student ETH Zurich), Karel Kubicek (ETH Zurich Information Security Group), Amit Zac (ETH Zurich Center for Law & Economics)

English Version: Dear Madam or Sir,

We contacted you [on [DATE]/ twice by email ([DATE1: SUBJECT1, DATE2 :SUBJECT2])] regarding research conducted at the Swiss Federal Institute of Technology in Zurich (ETH Zurich) about how websites comply with privacy laws according to the European General Data Protection Regulation (GDPR) and the EU ePrivacy Directive.

We would like to invite you to take part in our survey, which should only take about 5 minutes of your time. By filling out the questionnaire you are helping researchers and privacy regulators to understand the perspective and needs of websites that collect user data. Our aim is to better comprehend challenges businesses face regarding privacy laws.

Please find the survey and the Information form of this study at [LINK].

To give you further details about our study, we would like to inform you that we have chosen your website [URL] since our automated procedure observed that your website might not handle cookie notices and personal data collection correctly. We want to assure that there are no risks for you from this study:

We will not reveal the individual findings to anyone. We publish only aggregated results based on all 4000 websites in our study, and after this communication, we anonymize all our data to prevent unintentional leaks. We will not act against you legally. Our goal was to study the perspective of data collectors as you, and compare how helpful are information campaigns and self-assessment tools that we provide.

If you have any questions, please feel free to contact us via [COMMON EMAIL]. We thank you for your attention to this matter.

Sincerely,
Laura-Vanessa Soldner (Student ETH Zurich),
Karel Kubicek (ETH Zurich Information Security Group),
Amit Zac (ETH Zurich Center for Law & Economics)

Appendix B

Study Website

This is the content that is displayed on the Information Security Group website for the time of the study. The link to this website is be included in the notification, the reminder and the debriefing email.

Exploring Mechanisms of Website Compliance with Privacy Regulations

Overview We are currently conducting a scientific study to better understand how websites comply with privacy laws according to the European General Data Protection Regulation (GDPR) and the EU ePrivacy Directive. Specifically, we look at compliance in the area of cookie consent notices and cookie labelling.

The privacy regulations, including the General Data Protection Regulation and ePrivacy Directive, require websites to inform EU-based users about the collection of their data and request consent for the use of tracking technologies, such as non-essential cookies. We recognize the growing concern regarding the challenges faced by websites in maintaining compliance with privacy laws, especially concerning cookies and tracking technologies. In prior publication, we have shown that the majority of websites struggle to comply with privacy laws.

With our study, we would like to help website owners identify and fix security and data protection issues on their websites. To achieve this, we will be conducting a randomised control experiment where websites are contacted directly about this issue.

If you have any questions or would like to contact us regarding this research, please feel free to email us at cookie-compliance@inf.ethz.ch.

Publications Dino Bollinger, Karel Kubicek, Carlos Cotrini, David Basin. Automating Cookie Consent and GDPR Violation Detection. USENIX Security 2022 [4].

Appendix C

Survey

Information form

Boosting privacy compliance using self-auditing tools: The CookieAudit

Conducting person (full name): Laura-Vanessa Soldner

We would like to ask if your website is willing to participate in our research project. The participation is voluntary and 100% anonymous. Please read the text below carefully and ask us via email about anything you do not understand or would like to know. You can withdraw at any time, the survey will take 5 minutes of your time.

What is investigated and how?

Privacy regulations such as the General Data Protection Regulation and ePrivacy Directive (together 'privacy laws') require websites to inform EU-based users about the collection of their data and to request their consent to use tracking technologies such as non-essential cookies. There is growing concern that it is challenging for websites to stay compliant with privacy laws as far as cookies and tracking technologies are concerned.

We now ask for your help by filling out this short survey. The aim of this study is to investigate compliance with privacy laws. We are particularly interested in understanding the websites' perspective. This survey allows us to understand your motivation and problems in implementing privacy laws. This is a purely scientific study which will lead to a publication in a peer reviewed outlet. Results of the survey are collected and aggregated, reviewed only by the research team, and remain fully anonymised.

Who can participate?
You have to be over 18 years old and an employee or owner of the website. We randomly selected websites from our database of reviewed websites.

What am I supposed to do as a participant?
Answer a survey of 25 questions concerning the website and your institution.

What are my rights during participation?
Your participation in this study is voluntary. You may withdraw your participation at any time without specifying reasons and without any disadvantages.

What risks and benefits can I expect?
There are no specific risks to you or your website. As we explain below, we eliminate harm to your website reputation by securing your answers without any website identifiable information. Results of the survey are collected and aggregated, reviewed only by the research team, and remain fully anonymised.

Will I be compensated for participating?
No. We can offer you, free of charge, guidance on using our self-support tool to learn more about compliance with privacy laws, called CookieAudit. The tool can be downloaded here.

What data is collected from me and how is it used?

No personal data is collected on you. The answers to the survey are collected and saved in our dataset without your website contact details, URL or any other identifiers. Results of this study will only be published in an aggregated form, without any reference to a specific website. The anonymised data will not be shared outside ETH Zurich research team. Strict confidentiality will be observed at any time.

Who funds this study?
The study is funded by ETH Zurich and Swiss National Science Foundation.

Who reviewed this study?
This study was examined by the ETH Zurich Ethics Commission as proposal *EK-2023-N-55.*

Complaints office
The secretariat of the ETH Zurich Ethics Committee is available to help you with complaints in connection with your participation. Contact: *ethics@sl.ethz.ch* or 0041 44 632 85 72.

Survey Questions

1. Dear website representative. Thank you for filling out this anonymous survey. It will take approximately 5 minutes. Your participation supports a research project of the Swiss Federal Institute of Technology in Zurich (ETH). The survey is anonymous and follows data protection guidelines. Do you agree to include your website's answers in the study?

- Yes
- No

This survey focuses on the compliance with the regulations of the European General Data Protection Regulation (GDPR) and ePrivacy Directive, concerning cookie notices and personal data collection. From now on we refer to this as 'compliance with privacy laws' .

(NOT for control)
2. In (MONTH) this year`, we sent you the following notification. Did you receive this notification?

- Yes
- No
- Don't remember

3. What is your position in the organisation?

- We are a small business. We do not have a clear separation of roles.
- Operational level in IT (or similar)
- Management level in IT (or similar)
- Operational level in law (or similar)
- Management level law (or similar)
- Compliance officer
- Other

4. In which industry does your organisation operate?

- Information and communication
- Administrative and support service activities
- Transportation and storage
- Professional scientific and technical activities
- Water supply, sewerage, waste management and remediation activities
- Construction
- Business economy (except activities of holding companies)
- Wholesale and retail trade, repair of motor vehicles
- Mining and quarrying
- Manufacturing
- Real estate activities
- Accommodation and food service activities
- Other

5. How many employees does your organisation have (including yourself)?

- 1-2
- 3 to 9
- 10 to 49
- 50 to 249

- more than 249

6. Does your organisation employ an in-house legal advisor/lawyer(s)?
    - Yes
    - No
    - Don’t know

7. Who in your organisation is responsible for the maintenance of the website?
    - Single (IT) employee
    - IT department (or similar)
    - External company
    - Other

8. Who in your organisation is responsible for being compliant with the GDPR? [Mark all relevant answers]
    - CEO/Owner
    - No one, we use an External Company (e.g. Consent Management Platform)
    - Project management
    - Legal
    - IT department
    - PR department
    - Administrative department
    - Skip/No Answer

9. Does your organisation generate significant revenue by monetising data?
    - Yes, it is the core business
    - Yes, but it is not the core business
    - No
    - Don't know

10. Do you agree with the following statement: "Privacy is a core value of our organisation."?
    - [SLIDER] strongly agree, agree, disagree, strongly disagree OR skip

11. What could be the reason for a website to not comply with privacy laws? (allow several options)
    - Lack of technical knowledge on how to solve the problem
    - Time constraints
    - Problem has no priority
    - Problem was not known
    - Notification did not seem reliable

12. Do you agree with the following statement: "It is complicated to understand obligations from privacy laws."?
    - [SLIDER] strongly agree, agree, disagree, strongly disagree OR skip

13. Do you agree with the following statement: "It is complicated to implement services compliant with privacy laws"?
    - [SLIDER] strongly agree, agree, disagree, strongly disagree OR skip

(NOT for control)

14. How did your website approach becoming compliant with privacy laws?
    - Fixed it myself without help
    - Fixed it myself with help
    - Forwarded problem to colleague in organisation
    - Asked my external service provider to fix the problem
    - Hired a new service provider to fix the problem

15. Do you agree with the following statement: "Public authorities informed our organisation well enough to be compliant with privacy laws."?
    - [SLIDER] strongly agree, agree, disagree, strongly disagree OR skip

16. Do you agree with the following statement "Our compliance service provider, for example, cookie notice provider, informed our organisation well enough to be compliant with privacy laws."
    - [SLIDER] strongly agree, agree, disagree, strongly disagree OR skip

17. What do you estimate is the average number of websites which get fined each year because they are not GDPR compliant?
    - Less than 100
    - Between 100 and 399
    - Between 400 and 1000
    - Over 1000

18. Do you agree with the following statement: "Being compliant with privacy laws is a high cost for our organisation."?
    - [SLIDER] strongly agree, agree, disagree, strongly disagree OR skip

19. What is important for your choice of a technological solution to collect consent for the use of data collection, like cookie consent banner providers (CMPs)?
(each option has a [SLIDER] very important, important, slightly important, not important OR skip)
    - Price

- Collecting positive consent (i.e., accept all)
- Legal compliance
- Easy to set up
- Prior experience
- Other (open)

20. Do you agree with the following statement: "On the basis of the email notification I got from the research team, I could understand the legal requirements regarding cookie notices and personal data collection."?
    - [SLIDER] strongly agree, agree, disagree, strongly disagree OR skip

21. Do you agree with the following statement, if applicable: “The remainder email sent in [MONTH] was helpful for our organisation to become compliant.”
    - Yes, because _______ (open)
    - No, because ________(open)
    - No reminder email received

22. CookieAudit is a browser extension which targets web developers and data protection agencies (enforcers), allowing users to identify potential violations and informing them how to address these. Have you used CookieAudit in the past?
    - Yes
    - No

23. Do you agree with the following statement: "The CookieAudit tool was helpful to verify the compliance status of my website."?
    - [SLIDER] strongly agree, agree, disagree, strongly disagree OR skip

24. Do you agree with the findings of CookieAudit regarding your website?
    - Yes
    - No (open)

25. Do you agree with the following statement: "The CookieAudit tool was helpful to improve compliance with privacy laws on my website."?
    - [SLIDER] strongly agree, agree, disagree, strongly disagree OR skip

26. Is there anything else you would like to share?
    - (open)

Dear website representative. Thank you for your participation in this survey. For the privacy of website users, we would appreciate it if you consider (continue) to use and help us spread the word about the browser extension we have developed called **CookieAudit**.

CookieAudit can help website owners improve their compliance with privacy laws [link]. The extension allows you to identify potential privacy issues and informs you how to address these. It detects consent, used cookies, and reports potential privacy issues. CookieAudit is provided and operated by the Information Security Group at ETH Zurich [link]. It is free to use and there is no obligation to make any changes to your website based on the results.

ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Eigenständigkeitserklärung

Die unterzeichnete Eigenständigkeitserklärung ist Bestandteil jeder während des Studiums verfassten Semester-, Bachelor- und Master-Arbeit oder anderen Abschlussarbeit (auch der jeweils elektronischen Version).

Die Dozentinnen und Dozenten können auch für andere bei ihnen verfasste schriftliche Arbeiten eine Eigenständigkeitserklärung verlangen.

Ich bestätige, die vorliegende Arbeit selbständig und in eigenen Worten verfasst zu haben. Davon ausgenommen sind sprachliche und inhaltliche Korrekturvorschläge durch die Betreuer und Betreuerinnen der Arbeit.

Titel der Arbeit (in Druckschrift):

Quantifying Mechanisms behind Cookie Consent (Non-)Compliance: A Notification Study of Audit Tools

Verfasst von (in Druckschrift):

Bei Gruppenarbeiten sind die Namen aller Verfasserinnen und Verfasser erforderlich.

Name(n): Soldner

Vorname(n): Laura-Vanessa

Ich bestätige mit meiner Unterschrift:

- Ich habe keine im Merkblatt „Zitier-Knigge“ beschriebene Form des Plagiats begangen.
- Ich habe alle Methoden, Daten und Arbeitsabläufe wahrheitsgetreu dokumentiert.
- Ich habe keine Daten manipuliert.
- Ich habe alle Personen erwähnt, welche die Arbeit wesentlich unterstützt haben.

Ich nehme zur Kenntnis, dass die Arbeit mit elektronischen Hilfsmitteln auf Plagiate überprüft werden kann.

Ort, Datum

Zurich, 13.August 2023

Unterschrift(en)

Bei Gruppenarbeiten sind die Namen aller Verfasserinnen und Verfasser erforderlich. Durch die Unterschriften bürgen sie gemeinsam für den gesamten Inhalt dieser schriftlichen Arbeit.