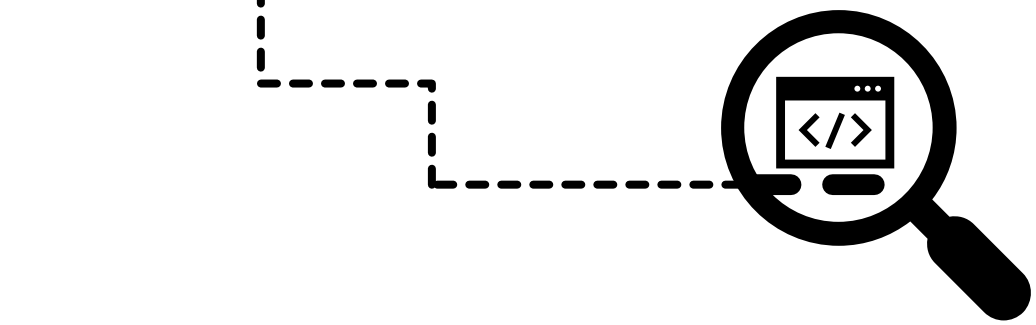
RUB

Third-Party Web Tracking Under the General Data Protection Regulation



Christine Utz

RUHR
UNIVERSITÄT
BOCHUM

RUB

THIRD-PARTY WEB TRACKING UNDER THE GENERAL DATA PROTECTION REGULATION

DISSERTATION

zur Erlangung des Grades eines Doktor-Ingenieurs
der Fakultät für Informatik
an der Ruhr-Universität Bochum

vorgelegt von

CHRISTINE UTZ

geboren in Augsburg

2023

Gutachter:	Prof. Dr. Thorsten Holz
	Prof. Dr. Florian Schaub
Tag der mündlichen Prüfung:	4. November 2022

IMPRINT

Christine Utz: *Third-Party Web Tracking Under the General Data Protection Regulation,* © 2023

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` (v4.6) developed by André Miede and Ivo Pletikosić. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both LaTeX and LyX:

`https://bitbucket.org/amiede/classicthesis/`

When algorithms judge everything we do,
we need to protect the right to make mistakes.

— Tijmen Schep

ABSTRACT

Modern websites make heavy use of third-party services for different purposes – to facilitate web development, connect to social media, or monitor visitors' interactions with the site to increase usability and revenue. Integration into a website allows the third party to collect personal information about the website's visitors, including their browsing history and implied interests and lifestyle. Unlike the first-party URL visible in the browser, the presence of third parties on a website and the associated data collection are often not apparent to visitors. Making data collection processes more transparent and giving people more control of the collection and use of their personal information were among the goals of the General Data Protection Regulation (GDPR) that went into effect in the European Union (EU) on May 25, 2018. Its possible implications for website owners have been widely discussed in the industry, motivating the development of third-party services that follow the GDPR's "privacy by default and by design" principle. Still, first research results could not identify a clear effect of the GDPR on the actual use of third-party services on websites.

This work investigates possible reasons and incentives for change by taking a closer look at the use and perception of third-party services on websites after the GDPR enforcement date, from the perspectives of both website owners and visitors. First, we explore if websites make increased use of transparency mechanisms regarding their data processing practices – instead of changing them. Though we find a notable increase in the prevalence of cookie consent notices after the GDPR enforcement date, they rarely allow visitors to provide free and informed consent to third-party data collection. To investigate in more detail how website visitors perceive and interact with cookie consent notices, we conduct a measurement study with different notice designs on a live website. If notices follow a transparent, privacy-by-default approach, people rarely opt into data collection by third-party services.

Turning towards the perspective of the people responsible for third-party use on websites, we conduct an online survey to investigate the underlying decision processes. We find a lack of awareness of the data collection through third-party services and that privacy only plays a role in vendor selection or configuration if there are legal obligations or guidelines. To raise awareness and incentivize change, we apply an approach used in the context of web security vulnerabilities and conduct a large-scale email notification campaign. For two months, we monitored security and privacy issues on websites, including the use of third-party cookies without visitors' consent. We find that website owners more positively respond to notifications about a security issue that is easy to fix, as opposed to privacy shortcomings that require more fundamental changes and are often not perceived as a problem.

Overall, this work identifies widespread misconceptions on both website visitors' and creators' sides about the privacy implications of third-party use on websites and available control mechanisms. Future work is encouraged to find ways to reduce the "consent burden" on website visitors, raise web developers' awareness of third-party data collection, and encourage the use of privacy-friendly alternatives.

KURZFASSUNG

Moderne Websites nutzen eine Vielzahl von Drittanbieterdiensten, etwa für effizientere Web-Entwicklung, zur Einbindung sozialer Medien oder zur Überwachung des Nutzungsverhaltens, um Usability und Umsatz zu steigern. Die Integration von Drittanbietern in eine Website erlaubt diesen Zugriff auf persönliche Daten der Besucher:innen, wie etwa ihren Browserverlauf und daraus abgeleitete Informationen über Lebensstil und Interessen. Während die Adresse der Website selbst im Browser sichtbar ist, sind die Existenz von Drittanbieterdiensten und die Datenerhebung durch diese für Nutzende der Website häufig nicht ohne Weiteres erkennbar. Datensammelprozesse transparenter zu machen und den Menschen mehr Kontrolle über Sammlung und Nutzung ihrer persönlichen Daten zu geben, gehört zu den Zielen der Datenschutzgrundverordnung (DSGVO), die in der Europäischen Union (EU) am 25. Mai 2018 in Kraft trat. Mögliche Auswirkungen der DSGVO auf Websites wurden im Vorfeld umfassend diskutiert, was die Entwicklung von explizit DSGVO-konformen, datensparsamen Drittanbieterdiensten befördert hat. Erste Forschungsergebnisse konnten dennoch keine klaren Auswirkungen der DSGVO auf den tatsächlichen Einsatz von Drittanbieterdiensten auf Websites feststellen.

Diese Forschungsarbeit untersucht mögliche Gründe und Anreize für Veränderung, speziell die Nutzung und Wahrnehmung von Drittanbieterdiensten auf Websites nach dem Inkrafttreten der DSGVO aus der Perspektive sowohl der Website-Betreibenden als auch der Besucher:innen. Eine erste Studie untersucht, ob Websites zunehmend Gebrauch von Transparenzmechanismen bezüglich ihrer Datenverarbeitungspraktiken machen, anstatt Letztere zu ändern. Obwohl nach dem Inkrafttreten der DSGVO deutliche Zuwächse in der Verbreitung von Cookie-Hinweisen zu verzeichnen sind, erlauben diese in den seltensten Fällen eine freie und informierte Einwilligung in die Datenverarbeitung durch Drittanbieterdienste. In einer Folgestudie mit verschiedenen Cookie-Hinweisen auf einer Live-Website wird untersucht, wie Website-Besucher:innen diese Hinweise wahrnehmen und mit ihnen interagieren. Sind die Hinweise bezüglich der Datenverarbeitung transparent und folgen dem Grundsatz "privacy by default", wird nur selten explizit einer Datensammlung durch Drittanbieterdienste zugestimmt.

Die Perspektive der Verantwortlichen für die Einbindung von Drittanbieterdiensten auf Websites wird mittels einer weiteren Online-Studie untersucht. Es zeigt sich ein Mangel an Problembewusstsein bezüglich der Datenerhebung durch Drittanbieterdienste. In den zugrunde liegenden Entscheidungsprozessen spielen Datenschutz und Privatheit bei Auswahl und Integration der Dienste nur eine Rolle, wenn es rechtliche Mindestanforderungen oder offizielle Richtlinien gibt. Um zu sensibilisieren und Anreize für Veränderung zu schaffen, wird ein Ansatz aus der Web-Sicherheitsforschung übernommen und eine groß angelegte E-Mail-Benachrichtigungskampagne durchgeführt. Zwei Monate lang werden Sicherheits- und Datenschutzprobleme auf Websites beobachtet, unter anderem die Verwendung von Drittanbieter-Cookies ohne vorherige Einwilligung. Es zeigt sich, dass die Betreiber:innen von Websites eher geneigt sind,

auf Benachrichtigungen wegen einfach zu behebender Sicherheitslücken zu reagieren, während Defizite im Bereich Datenschutz umfassendere Änderungen erfordern und häufiger nicht als Problem aufgefasst werden.

Insgesamt zeigen sich weit verbreitete Fehlvorstellungen über die Auswirkungen des Einsatzes von Drittanbieterdiensten auf Websites und verfügbare Mechanismen zur Steuerung ihrer Einbindung – sowohl auf Seiten derjenigen, die Websites betreiben, als auch der der Besucher:innen. Anknüpfungspunkte für künftige Forschung bieten sich in Möglichkeiten, die kognitive Last durch Zustimmungsdialoge zu reduzieren, das Problembewusstsein bezüglich Datenerhebung durch Drittanbieterdienste zu erhöhen oder zum Einsatz datensparsamer Alternativen zu motivieren.

ACKNOWLEDGMENTS

The last five years have been a wild ride with many ups (accepted papers! conferences! travel!) and also a few downs (COVID-19 lockdowns, paper rejections ...). I consider myself very lucky to have shared this journey with so many amazing people.

First, I would like to thank my advisor, Thorsten Holz, for giving me the freedom to work on the topics I was excited about, for providing such a great research and work environment, and for his always excellent advice on this PhD journey. Also many thanks to Florian Schaub for serving as the second reviewer for this thesis and for always providing great ideas and feedback in our joint research projects.

Further thanks go to Martin Degeling for bringing data protection and privacy topics to the SysSec group, for launching so many cool joint projects, and for being such a great person to share an office with.

This thesis would not exist without the awesome collaborators I had the honor to work with on many fun and challenging projects: Florian Farke, Theodor Schnitzler, Henry Hosseini, Sabrina Amft, Leonie Schaewitz, Franziska Herbert, Steffen Becker, Marvin Kowalewski, Markus Dürmuth, Thomas Hupperich, Sascha Fahl, Ben Stock, Matthias Michels, Ninja Marnau, Christopher Lentzsch, Tobias Urban, Martina Lindorfer, and Jakob Bleier. Also thanks to our student assistants Yana Koval, Ahmed Ali, and Mikka Rainer for supporting our projects with their hard work.

Many thanks also to the other great folks in the SysSec, MobSec, and InfSec groups, particularly Andre Pawlowski, Teemu Ryhtilahti, Dennis Tatang, Merlin Chlosta, Katharina Kohls, Maximilian Golla, Philipp Markert, and Ali Abbasi.

Further thanks go out to the fellow PhD students in my interdisciplinary graduate program, SecHuman – Security for People in Cyberspace: my tandem partner, Stephan Koloßa, the already mentioned Florian Farke and Steffen Becker, Jan Rensinghoff, Mary Shnayien, Benedikt Auerbach, Laura Kocksch, Alexander Helm, Olga Skrebec, Carina Wiesen, Benedikt Boenninghoff, Steffen Hessler, Marc Fyrbiak, and our wonderful coordinators, Astrid Wichmann, Susanne Kersten, and Anne Thiele.

Last, but certainly not least, I would like to thank Arne Reinsch for his immense support over the years even in the most stressful of times.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

A29DPWP	Article 29 Data Protection Working Party
ADPC	Advanced Data Protection Control
AIC	Akaike information criterion
ANOVA	Analysis of Variance
API	Application Programming Interface
BIC	Bayesian information criterion
CA	Certificate Authority
CAPTCHA	Completely Automated Public Turing test to tell Computers and Humans Apart
CCPA	California Consumer Privacy Act
CCC	California Civil Code
CDN	Content Delivery Network
CEO	Chief Executive Officer
CERT	Computer Emergency Response Team
CET	Central European Time
CIS	Commonwealth of Independent States
CISSP	Certified Information Systems Security Professional
CMP	Consent Management Platform
CMS	Content Management System
COVID-19	Coronavirus Disease 2019
CrUX	Chrome User Experience
CSP	Content Security Policy
CSS	Cascading Style Sheets
CV	Cramérs V
DDoS	Distributed Denial-of-Service
DKIM	DomainKeys Identified Mail
DMARC	Domain-based Message Authentication, Reporting and Conformance
DNS	Domain Name System
DNT	Do Not Track
DPA	Data Protection Authority
DPO	Data Protection Officer
EC	European Communities
ECJ	European Court of Justice

EDPB	European Data Protection Board
EU	European Union
GA	Google Analytics
GDPR	General Data Protection Regulation
GPC	Global Privacy Control
GPL	GNU General Public License
GPS	Global Positioning System
HQ	Headquarters
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
HTTPS	Hypertext Transfer Protocol Secure
IAB	Internet Advertising Bureau
IDE	Integrated Development Environment
IP	Internet Protocol
IPv6	Internet Protocol version 6
IRB	Institutional Review Board
MX	Mail Exchanger
NGO	Non-Governmental Organization
NIST	National Institute of Standards and Technology
NLP	Natural Language Processing
OBA	Online Behavioral Advertising
OpenWPM	Open Web Privacy Measurement
P3P	Platform for Privacy Preferences
PEM	Privacy-Enhanced Mail
RFC	Request for Comments
RTB	Real-Time Bidding
SD	Standard deviation
S/MIME	Secure/Multipurpose Internet Mail Extensions
SOP	Same-Origin Policy
SPF	Sender Policy Framework
SRI	Subresource Integrity
SSL	Secure Sockets Layer
SSO	Single Sign-On
SVN	Apache Subversion
TCF	Transparency and Consent Framework
TLD	Top-Level Domain
TLS	Transport Layer Security

UI	User Interface
UK	United Kingdom
URL	Uniform Resource Locator
US	United States
USA	United States of America
UX	User Experience
W3C	World Wide Web Consortium
XSS	Cross-Site Scripting

Part I

PROLOGUE

1 INTRODUCTION

1.1 MOTIVATION

Third-party resources on websites

Contemporary websites make frequent use of third-party services to integrate new functionality, design elements, or media resources. The underlying reasons are as multifaceted as the purposes for which external resources are used in web development. In today's Web, much content is monetized via online advertising and marketing [162], which frequently involves the inclusion of advertising networks to target ads to website visitors' presumed interests and web analytics to measure the success of online marketing campaigns. User expectations regarding the look and functionality of websites, paired with time and resource constraints in web development, also drive the adoption of third-party resources such as design frameworks, contact forms, and external media hosting [73].

Tracking of website visitors

This reliance on third parties can come at the cost of website visitors' data privacy. By embedding external resources, websites provide third-party vendors with the opportunity to collect personal data about the website's visitors, such as their IP address, device information, visited pages, and access to any long-term identifiers the third-party service may have stored in visitors' browsers [162]. High prevalence and mutual data sharing potentially allow third-party vendors to track people across the Web, to learn large shares of their browsing histories, and to use this information to infer interests, demographics, and lifestyle.

General Data Protection Regulation

These practices, though established for years [155, 288] and "quite pervasive" [73], may be at odds with privacy legislation. On May 25, 2018, the General Data Protection Regulation (GDPR) [78] went into effect in the European Union, with the goal to harmonize data protection laws across its member states. GDPR regulations demand that processing of personal data be rooted in one of six legal bases – including user consent –, be transparently communicated, and data processors follow a "data protection by design and by default" approach. The GDPR's implications for online businesses have been widely discussed in the months and years before it became enforceable, raising expectations of websites overhauling their data collection practices and using more privacy-friendly technology to comply with the new regulation.

Little observed change

Despite all this, the first web tracking measurements conducted after the GDPR went into effect have shown little change in the prevalence of user tracking though third-party services [162, 251] and even hint at a consolidation of data flows towards already powerful actors [103, 282].

The data collection through third-party services on websites to a large extent takes place without website visitors noticing [160], as the HTTP requests to third-party servers are not prominently shown in standard Web browsers and any information in the website's privacy policy about third-party use and data collection is unlikely to be read [202]. Thus, the observed lack of change in websites' practices in third-party web tracking after the GDPR enforcement date directly contradicts the regulation's designated goals to introduce a consistently

high standard of prerequisites for the collection of personal data in Europe and provide users with individual rights regarding how companies handle their personal data. Presumed reasons for this lack of change in observed tracking practices include a lack of enforcement and not enough regulatory guidance concerning the use of concrete technology, though multiple decisions by national courts and data protection authorities in early 2022 [22, 149, 198, 217] indicate some positive development in this area, with more similar decisions expected to follow [26]. The use of third-party resources on websites could also become the subject of future legislative developments that directly target third-party vendors: As the most popular third-party services are offered by large Internet companies including Google and Facebook, the use of their services on websites contributes to centralization of information and, thus, power with these companies, which is currently under scrutiny of regulators both in the European Union (EU) and the United States (US) [37, 97].

Regulatory deficits

Research gaps

Previous work has shown that when Internet users are made aware of the personal information third-party services on websites can collect and infer about them, they often express surprise or even shock at the amount of data or number of services involved [274, 292]. But even the people responsible for the use of third-party services by websites – web developers, operators, and people in related roles – may be unaware of the extent of this data collection. As web tracking research has shown, many third-party services issue requests to other third parties, whose identity is not always deterministic, particularly in the context of real-time auctions of websites' advertising space [280]. This results in a complex web of interconnected entities that can get access to website visitors' personal information [118]. Website creators' lack of awareness and consideration of the privacy implications of third-party use are one potential reason for the apparent lack of change in websites' third-party tracking practices that had been largely unexplored.

What could be observed though in the months leading up to the GDPR enforcement date was anecdotal evidence hinting at an increase in websites' use of *(cookie) consent notices*, banners or prompts shown to users upon first visit that ask to acknowledge the website's use of cookies or similar tracking technologies. This suggests that websites reacted to the GDPR by means of increased transparency, rather than an actual change in data processing practices. While prior work had studied privacy policies as the primary mechanism to inform website visitors about a website's data processing practices, there were only limited insights into the prevalence of consent notices, their implementation, or visitors' perception of them.

Research goals

This thesis looks beyond the measured lack of change in third-party tracking practices after the GDPR enforcement date and investigates possible underlying reasons, other effects of the GDPR on websites, and opportunities to incentivize compliance. More concretely, we explore the hypothesis that websites primarily reacted to the GDPR by means of increased transparency about their data processing practices, rather than changing them, and investigate the prevalence, implementation, and user perception of cookie consent notices. In order to gain further insights into the lack of effect of the GDPR on third-party use on websites despite its "data protection by design" principle, we study website oper-

ators' considerations in the integration of third-party services. We also explore one promising method to help website operators comply with privacy legislation: large-scale automatic detection of complex privacy issues and notifying operators about them.

Our findings inform lawmakers, data protection authorities, and the research community about widespread misconceptions regarding the data collection through third parties on websites and the mechanisms offered for its control. We identify opportunities for future work to incentivize compliance through standardization of web privacy mechanisms and the development of guidelines that help web operators and software vendors create websites and consent mechanisms that follow the GDPR's "data protection by design and by default" principle.

1.2 TOPIC AND CONTRIBUTIONS

This thesis investigates third-party web tracking under the GDPR from two perspectives: website visitors, whose personal information can be collected and processed by third-party services on websites (Part ii, Chapters 3 and 4), and the people responsible for the integration of third-party services into websites (Part iii, Chapters 5 and 6). While the first three of these chapters are concerned with determining the status quo, the fourth investigates one possible approach towards improvement of website visitors' privacy. In the following, we provide an overview of the topics investigated in these chapters, what prior work has done in the area, and our contributions.

1.2.1 *Prevalence and Implementation of Consent Notices*

Traditionally, privacy policies have been the main transparency mechanism to provide information about the data processing practices of a service or business, including websites. Over the last decade, harmonization efforts in European privacy law have introduced the requirement for websites to ask their visitors for consent before storing data on their visitors' terminal equipment for purposes beyond what is strictly necessary to provide the requested service. This has prompted websites targeting people in Europe to also make privacy information available via *(cookie) consent notices*, colloquially referred to as "cookie banners." These are dialogs shown by websites upon first access that request visitors to acknowledge the website's use of cookies or similar tracking technologies. Originating in a 2009 addendum to the European Union's ePrivacy Directive [77] that became mandatory for EU member states to implement into national laws by 2011, they initially were not widely used, as many member states had not timely implemented the directive [20]. Only a few years later, consent notices appeared to become more widespread when businesses started to prepare for the General Data Protection Regulation (GDPR) to come into effect on May 25, 2018. The goal of this law, directly applicable in all member states, was to harmonize data protection standards across Europe. It introduced six legal bases for data processing, including the data subject's consent. As a result, websites appeared to increasingly resort to consent notices to also ask for consent under the GDPR. At the same time, the options these notices

provide to website visitors also started to become more complex, including the ability to provide or deny consent individually for each third-party service used by a website. While there is a rich body of work on the prevalence [63], mechanisms [226], and evolution [155, 288] of web tracking through third parties, prior to 2018 there had not been the opportunity to study changes due to new privacy legislation that is expected to have a direct effect on how websites process and protect user data and inform visitors about these practices. Research investigating websites' transparency and control mechanisms had focused on privacy policies [110, 194] and opt-out mechanisms for targeted advertising [90, 154], while there were only limited insights into the prevalence [13] and user perceptions [147] of consent notices. Thus, the aforementioned development in the use of consent notices in the months leading up to the GDPR enforcement date gave rise to the first research question this thesis investigates:

RQ 1: Has the GDPR provided website visitors with greater transparency and control regarding the collection of their personal information by third-party services on websites?

Measurement study of transparency mechanisms

To answer this question, in Chapter 3 we conduct a longitudinal measurement study on popular websites across all 28 EU member states[1] and measure the prevalence of third-party cookies and consent notices in the months before and after the GDPR enforcement date. We classify consent notices by the granularity they offer in their interfaces to accept or deny data collection, including the possibility to do this for distinct categories of third-party functionality or for each third party individually. Moving beyond the interface, we investigate popular third-party cookie consent libraries for whether their backend actually allows for an implementation that honors website visitors' decisions. Between January and the end of May 2018, we find a 16 % higher prevalence of consent notices as a transparency mechanism, but only limited increase in control, as less than 20 % of websites offer visitors the opportunity to deny data collection in consent notice interfaces and popular consent libraries frequently lack a backend capable of honoring the visitor's choice.

1.2.2 *Website Visitors' Perception of Consent Notices*

This increase in consent notices' prevalence and complexity, paired with a lack of underlying functionality to implement user choice, left us with the impression that the landscape of consent notices shortly after the GDPR enforcement date did not provide website visitors with meaningful options to control data collection on websites. This sentiment was apparently shared by many other Web users, as evidenced by reports of consent fatigue [25], as well as the emergence of browser plugins that prevent consent notices from being displayed [143].

As many websites defaulted to data collection or did not offer any control mechanisms, this development contradicted the GDPR's and ePrivacy Directive's intentions and regulations. This made us wonder how, in contrast to existing practice, website visitors perceived and interacted with consent notices that made better use of the available design space and provided them with actual

1 As of 2018, when the United Kingdom was still a member of the EU.

choices to control a website's data processing practices, including data collection through third parties. Thoroughly exploring alternative options for the design of consent notices' user interfaces can inform regulators looking into providing more concrete guidance on the implementation of consent to help websites comply with the law. It can also provide actors in the web tracking ecosystem with insights about the consent rates to expect from users in case regulators enforce consent requirements at larger scale.

Prior work has partially identified parameters of the design space of consent notices' user interfaces, including the amount of information they provide [147], options for interaction [234], the latter of which we also investigate in Chapter 3, and blocking behavior [234]. Others collected metrics about notices' size, location, and word, link, and button counts [284]. While this was mostly done in the context of measurement studies, the work of Kulyk et al. [147] studied user perceptions of consent notices. As the focus of this study lay on different wordings, it did not examine the interaction options available to website visitors or other parameters of consent notices' user interfaces.

Thus, our work in Chapter 4 addresses this research gap and investigates the following question:

RQ2: How do website visitors perceive and interact with different types of consent notices, if given an actual choice to allow or deny consent?

Field study of consent notices

We had the opportunity to evaluate different consent notice designs on a live website, while the aforementioned user study was conducted in a lab setting. Thus, we investigate this research question by conducting the first study of cookie consent notices in a real-world environment. We use a sample of 1,000 notices from live websites to determine the design space for the user interface of consent notices, identifying eight parameters. In an iterative study design, we investigate the effect of the following user interface (UI) parameters on visitors' interactions: location on the site, available options and use of nudging, presence of a privacy policy link, and use of technical vs. non-technical language.

We find that website visitors are most likely to interact with consent notices displayed in the bottom left corner of the viewport and offer a binary choice, while technical language or a privacy policy link do not have any notable influence on interaction rates. Nudging towards accepting all cookies significantly impacts visitor interaction with consent notices, particularly preselections in notices that offer a fine-grained selection based on categories or third-party vendors: If options are preselected, about 10 % of visitors on desktops and 30 % of mobile users accept data collection through all categories or vendors, while only around 4 % allow data collection through some vendors and less than 0.1 % allow all data collection if checkboxes are not preselected. Incorrect mental models about the workings of consent notices are widespread, including the assumption that the website cannot be accessed unless consent is given. This can be partly due to previous experience with websites that had implemented consent incorrectly. Despite these misconceptions, our results indicate that the vast majority of website visitors, when presented with consent notices that correctly implement freely given, specific, informed, and unambiguous consent as demanded by European privacy law, are unlikely to consent to data collection by third-party services.

1.2.3 *Websites' Considerations in the Use of Third-Party Services*

Consent notices that negatively affect user experience would not be necessary if websites did not collect any visitor data subject to the consent requirement, which is often triggered through use of third-party services [11, 135]. Thus, Part iii shifts the perspective towards the people responsible for the use of third-party services on websites – web developers and people working on websites in related roles, such as administrators, content creators, or social media managers. As it is them who add third-party services to websites, thus enabling them to collect visitors' personal data, one can argue that it is their responsibility to ensure that visitors' privacy is considered in the design and implementation of a website [243] – and even more so under the GDPR's "privacy by design and by default" mandate.

The question whether user privacy is considered in the software development process has been investigated in different contexts. Research that specifically investigates developers' use of third-party resources exists for smartphone applications [215, 232, 233], and privacy considerations in particular have been explored for the selection and configuration of mobile ad networks [183, 266]. By contrast, little is known about the decision processes that lead to the use of third-party services on websites, for which different mental models regarding their data processing practices might apply. Thus, in Chapter 5 we pose the following research question:

RQ3: Do people working with websites consider visitors' privacy in the integration of website functionality that is often integrated via third-party services?

Online study with website creators

To answer it, we conduct the first online study on websites' privacy practices in their use of third-party services. We combine survey answers with web measurements to learn about the prevalence of first- vs. third-party integrations for ten common types of website functionality, the decisions that led to the adoption of these integrations, and whether visitor privacy was considered in the process. We also study use of privacy-enhancing configurations in integration and website creators' awareness of the data collected through third parties. We find websites' use of third-party services to differ across different types of functionality. Most participants reported not to have looked into alternatives to the chosen solution, but those who did mostly considered a first-party solution, rather than another third-party service. The main factors for third-party adoption are ease of integration and familiarity with an existing service. By contrast, visitors' privacy only influences the decision process when guidelines by data protection authorities (DPAs) or court rulings provide concrete guidance, as in the case of web analytics [145]. A potential reason are widespread misconceptions of the privacy implications of third party use: While participants seem to be aware of data collection directly associated with the purpose for which the third-party service is used, this does not extend to less prominent data flows such as the transfer of IP addresses and device information.

1.2.4 *Email Notifications about Third-Party Web Tracking Without Consent*

In the light of this assessment, the natural follow-up question is what could be done to raise website owners' awareness of the potential privacy implications of the use of third-party services and incentivize compliance with privacy legislation. One approach previously used in security and privacy research to alert website owners about diverse issues with their websites are large-scale *notification studies*. As outlined by Maass et al. [168], this line of research typically involves identifying a set of hosts that are affected by a specific security vulnerability or compliance issue of interest, establishing a channel of communication, and notifying them about the issue. Afterwards, the hosts are periodically rechecked to determine if and when the issue has been fixed. Additional, direct feedback is frequently collected through channels such as self-service tools, email correspondence, or surveys.

Starting about a decade ago, such notification studies were first conducted to alert website operators about abuse of their infrastructure for malicious purposes, including distribution of drive-by downloads [286] and URLs serving malicious downloads from botnets [33], which resulted in higher fix rates compared to unnotified affected systems. This approach was subsequently adopted to notify websites about different web security vulnerabilities as diverse as Heartbleed [59], HTTPS misconfigurations [299], DDoS amplifiers [146], Cross-Site Scripting (XSS) [261], or accidental leakage of sensitive information in repositories of software versioning systems [167, 260]. Most recently, notification campaigns have also been conducted to alert website owners of privacy issues, though limited to specific third-party services: use of Google Analytics without IP anonymization [169] and incorrect implementations of the OneTrust Consent Management Platform (CMP) that do not constitute valid consent to data processing under the GDPR [197]. This focus on specific providers for privacy notifications reflects one of the core challenges of any notification study: the requirement to identify the investigated issue(s) remotely, without the need for manual verification, and with as high as possible accuracy. This is because the goal of any ethical and responsible notification study is to avoid false positives to not cause unnecessary anxiety and cost on the recipients' side. In the case of privacy issues, this means that the issue at hand must constitute a clear violation of unambiguous requirements mandated by applicable data protection law, which can be hard to determine without human verification. This gave rise to the following research question:

RQ4: Can email notifications motivate website operators to fix complex privacy issues, including use of third-party cookies without visitors' consent?

Large-scale email notification study

In Chapter 6 we tackle this challenge and investigate the feasibility of large-scale email notification campaigns for more complex, vendor-independent privacy issues, including two related to third-party web tracking: use of third-party cookies (i) without a consent notice and (ii) in presence of a notice but before the visitor has consented to their use. We compare fix rates and feedback from survey responses and email communication with the performance of notifications about a potential security vulnerability. Though we observe only limited influence of our notifications on remediation rates, we find that email

notifications about privacy issues are not as well perceived as notifications about a potential security vulnerability. They result in lower fix rates, less incentive to take action, and more negative feedback. Investigations for possible reasons confirm website operators' lack of awareness of third-party data collection capabilities identified in Chapter 5 and find additional misconceptions regarding the presumed inapplicability of privacy legislation.

1.3 OUTLINE

The rest of this thesis is organized as follows: In Chapter 2, we introduce the role of third-party services in web development, the implications of their use on website visitors' data privacy, and possible remediations. We also provide an overview of the GDPR provisions that may impact the use of third parties on websites and briefly introduce other privacy legislation that plays a role in subsequent chapters, namely the EU's ePrivacy Directive and the California Consumer Privacy Act (CCPA).

Part ii investigates the GDPR's impact on third-party use from the perspective of website visitors. For them, the regulation coming into effect has resulted in a visible increase in websites using (cookie) consent notices. In Chapter 3 we study their prevalence before and after the GDPR enforcement date and identify to what degree they allow website visitors to control the collection of their personal data through third-party services. Observing an overall lack of meaningful choices, in Chapter 4 we study how people interact with different types of consent notices, including those that comply with privacy law, and investigate if people actually consent to third-party data collection if given the choice.

Part iii switches the perspective to the people responsible for the integration of third-party services into websites – web developers and people in related roles. In Chapter 5 we investigate their decision processes behind the selection and configuration of third-party services for different popular types of website functionality and explore if they are aware of the privacy implications of third-party use. Observing a lack of awareness of data collection through third parties, in Chapter 6 we conduct a large-scale email notification study to alert website owners of a variety of privacy issues including use of third-party cookies without consent and monitor if these notifications can motivate remediation.

From our results, we identify opportunities for future work in Chapter 7 and conclude with a recapitulation of our findings and contribution in Chapter 8.

1.4 LIST OF PUBLICATIONS

Publications that serve as the basis of this thesis

The basis for this thesis are four peer-reviewed papers that have either already been published or are currently under submission. They are listed in the following, along with individual contributions.

1. Martin Degeling, Christine Utz, Henry Hosseini, Christopher Lentzsch, Florian Schaub, and Thorsten Holz: "We Value Your Privacy ... Now Take Some Cookies: Measuring the GDPR's Impact on Web Privacy." In: *Proceedings of the 2019 Network and Distributed System Security Symposium (NDSS '19)*. San Diego, CA, USA: Internet Society, 2019. DOI: 10.14722/ndss.2019.23378, URL: https://www.ndss-symposium.org/ndss-paper/we-value-your-privacy-now-take-some-cookies-measuring-the-gdprs-impact-on-web-privacy/.

   For this work, I developed the classification of cookie consent notices and analyzed the technical features of popular consent libraries. I also performed manual annotation of websites for privacy policies and consent notices, as did everyone among the first four on the author list. Martin Degeling had the idea for this project, created the website crawler and web annotation tool, and analyzed the prevalence of privacy policies and consent notices. Henry Hosseini performed text analysis of the collected privacy policies; the results of the analyses related to privacy policies will be included in his thesis. Christopher Lentzsch conceptualized and computed the score measuring the reach of specific consent libraries.

2. Christine Utz, Martin Degeling, Sascha Fahl, Florian Schaub, and Thorsten Holz. "(Un)informed Consent: Studying GDPR Consent Notices in the Field." In: *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (CCS '19)*. London, United Kingdom: ACM, 2019, pp. 973–990. DOI: 10.1145/3319535.3354212. URL: https://dl.acm.org/doi/10.1145/3319535.3354212.

   For this follow-up work I identified design parameters for the user interface of consent notices, devised the notice designs, programmed their front end, implemented the survey, and analyzed the responses. Martin Degeling arranged the study logistics, implemented the consent notices' backend, and analyzed the measured user interactions.

3. Christine Utz, Sabrina Amft, Martin Degeling, Thorsten Holz, Sascha Fahl, and Florian Schaub: "Privacy Rarely Considered: Exploring Considerations in the Adoption of Third-Party Services by Websites." In: *Proceedings on Privacy Enhancing Technologies* 2023.1 (January 2023), pp. 5–28. DOI: 10.56553/popets-2023-0002. URL: https://petsymposium.org/popets/2023/popets-2023-0002.php.

   I came up with the original idea for this work, created the keyword lists for website-based recruitment, designed, implemented, and distributed the survey, analyzed the results, and wrote the majority of the paper. Sabrina Amft collected and processed email addresses for GitHub-based recruitment, helped with coding of the open-ended answers, and compiled related

work. Martin Degeling conducted the OpenWPM crawls for website-based recruitment and analysis of the websites provided by participants.

4. Christine Utz, Matthias Michels, Martin Degeling, Ninja Marnau, and Ben Stock: "Comparing Large-Scale Privacy and Security Notifications." In: *Proceedings on Privacy Enhancing Technologies* 2023.3 (May 2023), pp. 173–193. DOI: 10.56553/popets-2023-0076. URL: https://petsymposium.org/popets/2023/popets-2023-0076.php.

   For this project I implemented the survey, initiated and monitored the daily privacy checks, analyzed the resulting measurement data and the closed-ended survey responses, helped with classification of email correspondence, and wrote large parts of the paper. Martin Degeling and Ahmed Ali implemented the OpenWPM-based privacy checks. Matthias Michels built the notification infrastructure, implemented and ran the Git checks, and conducted and classified all communication with notification recipients. Ninja Marnau provided legal advice and analyzed the open-ended survey questions. Charlotte Schwedes and Simon Lenau of CISPA Empirical Research Services computed the logistic regressions.

Other publications

During the time of this dissertation, there was opportunity to contribute to other peer-reviewed publications outlined in the following that are not considered in this thesis.

1. Theodor Schnitzler, Christine Utz, Florian Farke, Christina Pöpper, and Markus Dürmuth. "User Perception and Expectations on Deleting Instant Messages – or – 'What Happens If I Press This Button?'" In: *Proceedings of the 3rd European Workshop on Usable Security (EuroUSEC 2018)*. London, United Kingdom: Internet Society, 2018. DOI: 10.14722/eurousec.2018.23009. URL: https://www.ndss-symposium.org/wp-content/uploads/2018/06/eurousec2018_09_Schnitzler_paper.pdf.

   Here I helped conduct the actual study in Ruhr University Bochum's dining hall and was involved in data analysis and paper writing. The project lead was Theodor Schnitzler, who created the study design together with Florian Farke.

2. Theodor Schnitzler, Christine Utz, Florian Farke, Christina Pöpper, and Markus Dürmuth. "Poster: User Perception and Expectations on Deleting Instant Messages – or – 'What Happens If I Press This Button?'" In: *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*. Baltimore, MD, USA: USENIX Association, 2018.

   I provided feedback on this poster created by Theodor Schnitzler about our EuroUSEC 2018 paper. Together we presented it at the SOUPS 2018 poster session.

3. Christine Utz, Stephan Koloßa, Thorsten Holz, and Pierre Thielbörger: "Die DSGVO als internationales Vorbild?" In German; English title: "The GDPR as an International Role Model?" In: *Datenschutz und Datensicherheit* 43 (Nov. 2019), pp. 700–705. DOI: 10.1007/s11623-019-1192-5.

URL: https://link.springer.com/article/10.1007%2Fs11623-019-1192-5.

This article was part of a special issue featuring the members of my interdisciplinary graduate program, SecHuman – Security for People in Cyberspace. I wrote the majority of this article and Stephan Koloßa provided the parts about his research in international law.

4. Theodor Schnitzler, Christine Utz, Florian Farke, Christina Pöpper, and Markus Dürmuth. "Exploring User Perceptions of Deletion in Mobile Instant Messaging Applications." In: *Journal of Cybersecurity* 6.1 (2020), pp. 1–15. DOI: 10.1093/cybsec/tyz016. URL: https://academic.oup.com/cybersecurity/article/6/1/tyz016/5718217.

   This is an extended version of our EuroUSEC 2018 paper with additional analyses. I contributed to writing and editing.

5. Christine Utz, Steffen Becker, Theodor Schnitzler, Florian M. Farke, Franziska Herbert, Leonie Schaewitz, Martin Degeling, and Markus Dürmuth. "Apps Against the Spread: Privacy Implications and User Acceptance of COVID-19-Related Smartphone Apps on Three Continents." In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI 2021)*. Yokohama, Japan: ACM, 2021, pp. 1–22. DOI: 10.1145/3411764.3445517. URL: https://dl.acm.org/doi/10.1145/3411764.3445517.

   The idea for this longitudinal study of people's perception of smartphone applications created to help fight the COVID-19 pandemic emerged during the first COVID-19 wave in spring 2020. I helped analyze real-world apps to create the vignettes used in the study, implemented a test survey, performed qualitative data analysis with Leonie Schaewitz, computed inter-coder reliability, and oversaw the writing process. Everyone on the author list was involved to some degree in the study design, creation of the survey, data analysis, and writing. Martin Degeling computed the statistical models.

6. Henry Hosseini, Martin Degeling, Christine Utz, and Thomas Hupperich. "Unifying Privacy Policy Detection." In: *Proceedings on Privacy Enhancing Technologies* 2021.4 (July 2021), pp. 480–499. DOI: 10.2478/popets-2021-0081. URL: https://petsymposium.org/popets/2021/popets-2021-0081.php.

   For this paper I helped with brainstorming and paper writing. Lead author on this project was Henry Hosseini, who did the implementation and evaluation.

7. Martin Degeling, Christine Utz, Florian M. Farke, Franziska Herbert, Leonie Schaewitz, Marvin Kowalewski, Steffen Becker, Theodor Schnitzler, and Markus Dürmuth. "Die Nutzung von Smartphone-Apps zur Eindämmung von COVID-19 in Deutschland." In: *Technologien der Krise – Die COVID-19-Pandemie als Katalysator neuer Formen der Vernetzung*. Ed. by Dennis Krämer, Joschka Haltaufderheide, and Jochen Vollmann. Transcript Verlag, July 2022. ISBN: 978-3-8376-5924-5. URL:

https://www.transcript-verlag.de/978-3-8376-5924-5/technologien-der-krise/.

This article presents selected findings from our longitudinal study of COVID-19 apps. I contributed to writing and editing.

8. Siddhant Arora, Henry Hosseini, Christine Utz, Vinayshekhar Bannihatti Kumar, Tristan Dhellemmes, Abhilasha Ravichander, Peter Story, Jasmine Mangat, Rex Chen, Martin Degeling, Tom Norton, Thomas Hupperich, Shomir Wilson, and Norman Sadeh. "A Tale of Two Regulatory Regimes: Creation and Analysis of a Bilingual Privacy Policy Corpus". In: *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*. Marseille, France: European Language Resources Association (ELRA), 2022, pp. 5460–5472. URL: http://www.lrec-conf.org/proceedings/lrec2022/pdf/2022.lrec-1.585.pdf.

   This paper is the first in a project with Carnegie Mellon University that aims to collect, annotate, and analyze a bilingual corpus of privacy policies in English and German. I was involved in the selection of apps with privacy policies in both languages, comparison of privacy policy texts across languages, and classification of differences, as was Henry Hosseini. A changing group of people at CMU created the machine learning classifiers, the current lead being Siddhant Arora. Martin Degeling organized and monitored the hiring of annotators for the German privacy policies.

9. Christine Utz, Steffen Becker, Theodor Schnitzler, Florian M. Farke, Franziska Herbert, Leonie Schaewitz, Martin Degeling, and Markus Dürmuth. "Poster: Apps Against the Spread: Privacy Implications and User Acceptance of COVID-19-Related Smartphone Apps on Three Continents." In: *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*. Boston, Massachusetts, USA: USENIX Association, August 2022. URL: https://www.usenix.org/conference/soups2022/presentation/utz-poster.

   I designed and created this poster about our first COVID-19 app paper published at CHI 2021.

10. Marvin Kowalewski, Christine Utz, Martin Degeling, Theodor Schnitzler, Franziska Herbert, Leonie Schaewitz, Florian M. Farke, Steffen Becker, and Markus Dürmuth. "52 Weeks Later: Attitudes Towards COVID-19 Apps for Different Purposes Over Time." To be published at the 26th ACM Conference on Computer-Supported Cooperative Work and Social Computing, October 2023.

    This is the second conference paper in our longitudinal project about COVID-19 apps. As in the first paper published at CHI 2021, I conducted the qualitative data analysis with Leonie Schaewitz, computed inter-coder reliability, and contributed to paper writing.

2 BACKGROUND

As a foundation for this thesis, this chapter introduces the concept of third-party services in web development – their advantages, disadvantages, common use cases, privacy implications for website visitors, and ways to remediate them. We also provide an overview of the relevant privacy legislation that impacts their integration into websites, particularly the General Data Protection Regulation (GDPR) as the focus of this work.

2.1 THIRD-PARTY SERVICES IN WEB DEVELOPMENT

2.1.1 *Overview*

Definition

In the context of websites, third parties are "entit[ies] outside the primary site–user relationship, i. e., the aspects of the site not directly within the control of the site owner but present with their approval" [121]. More concretely, if a website `https://institution.org` (the *first party*) is accessed by a visitor (the *second party*), its code and content, such as HTML files, JavaScript code, images, videos, or other media files may be stored on the host system itself (`https://institution.org`) or retrieved from an another host operated by a different entity (the *third party*), such as `https://videohoster.com`.

Purposes

These remote resources can range from simple outsourced file storage to decrease resource use on the first-party host system (e. g., media hosting, content distribution networks) to the inclusion of complex software into the website for a myriad of different purposes from design frameworks to web analytics and advertising. In Section 2.1.5 we investigate common use cases for third-party functionality from the website perspective in more detail.

Actors

Perspective matters, as the involved actors' interests in third-party services differ: A third-party service deemed useful by website owners to provide a specific functionality may offer that functionality but, as we will see shortly in Section 2.1.3, may also be used by the third-party vendor to collect data about the website's visitors for advertising purposes [73, 162]. For instance, Libert and Nielsen [162] name the illustrative example of a social media "share" button: It provides website owners with the functionality to help visitors share the website's content on social media, while the respective social media service might be mainly motivated by this inclusion into websites allowing the social media company to track what websites their users visit, regardless of whether an individual actually clicks the share button [162].

Prevalence

Over the last two decades, websites have come to rely on an increasing number of third-party services, as shown by longitudinal studies performed on archival data [155, 288]. For example, Lerner et al. found that the percentage of websites that issued requests to at least five third-party domains had risen from about 5 % in the early 2000s to about 40 % in 2016 [155]. Websites of public institutions tend to use fewer third parties than those of private compa-

nies, with news websites featuring the highest number of requests to distinct companies [251].

Similarly, the coverage of third-party services has increased over time: Measurements on archival data provided evidence that in the early 2000s no single third-party tracking service was present on more than 10 % of examined websites [155, 288]. As of the late 2010s, the third-party web tracking landscape is governed by a "long tail" distribution, with thousands of distinct observed third-party domains, most of them used for advertising [251], but only a small number of them present on a significant share of websites [63, 251] and centralization continuing to increase [103, 282]. For years, the majority of the most prevalent third-party domains have been domains owned by Google [63, 92, 139, 160, 251], including those used for Google Analytics (`google-analytics.com`), which is the most widely used third-party service on the Web with reported coverages of 70 % and higher [63, 92]; Google's advertising networks including DoubleClick (`doubleclick.com`); Google Fonts (`fonts.google.com`); and domains used for content delivery including `gstatic.com` and `google.com`.

2.1.2 *Benefits of Third-Party Use in Web Development*

This increase in third-party use and coverage reflects the evolution of the Web from static pages to complex applications running in visitors' browsers. In the light of this, web development increasingly faces the challenge to reconcile increased user expectations of a web service's design and functionality with limited human and monetary resources [73], thus fueling the need to reuse existing code and design resources. While other mechanisms for code reuse in web development do exist, as we will see shortly in Section 2.1.4, particularly widespread is the inclusion of third-party code remotely hosted on another server. For web developers, probably the most important reported benefit to use remotely hosted third-party resources is ease of integration – often all that is required is to copy and paste a short HTML or JavaScript snippet from the vendor's website [225]. Outsourcing functionality also moves responsibility for the maintenance or development of the functionality to the third party. Many popular third-party services are available free of charge – which might be one of the main reasons for their widespread use. Further, remote resources are perceived to speed up website load times if they are served by Content Delivery Networks (CDNs) or widely used and thus likely to be cached in the visitor's browser [216, 225]. Third-party domains have also been used to work around browser limits for open connections to a single domain [216].

2.1.3 *Risks of Third-Party Use on Websites*

Non-privacy risks

On the other hand, the inclusion of external JavaScript resources can negatively affect website load times due to network negotiation overhead [122, 216, 225] and introduces dependencies on the availability of third-party code [216]. The potentially far-reaching implications of such dependencies were demonstrated in 2016, when the delisting of a simple string padding function from `npm`, a JavaScript package manager, broke numerous websites across the globe [41]. Third parties can also introduce security risks: A lack of HTTPS support by

third-party vendors has hindered adoption by the websites that use them [63], and libraries that are no longer maintained or not updated can leave the website vulnerable to known attack vectors [151]. Remotely delivered content could be replaced with malicious payload, for example, by compromising the legitimate third party or via typosquatting attacks [193]. While countermeasures such as Subresource Integrity (SRI) do exist, they only apply to static assets [216, 259]. Third-party use can also interfere with other web security mechanisms, for example, prompting web developers to use more lenient policies in the deployment of Content Security Policy (CSP), a countermeasure against Cross-Site Scripting (XSS) [227, 259].

Generic privacy risks

Further, the seemingly harmless retrieval of resources from third-party servers can have implications for the data privacy of the website's visitors, which is the focus of this work. Due to the properties of the HTTP protocol, requests to a third-party resource include the URL of the embedding web page in the `Referer` [sic] header, as well as the visitor's IP address and their `User-Agent` string, which can contain information about their browser, operating system, and device [160]. Under EU law, IP addresses are considered personal data [67], as they allow for (if only temporary) identification of an individual.

If a third-party vendor is present on a large number of websites, the referrer and device information retrieved via repeated HTTP requests from different websites can allow them to track Web users across sites and learn the associated parts of their browsing history [160]. This is often accompanied by the direct request allowing the third party to place its own cookies and other long-term identifiers in visitors' browsers. Along with stateless tracking techniques such as browser fingerprinting this makes user reconnaissance across websites and browsing sessions even easier, though IP addresses alone were shown to still have significant tracking capabilities [184].

The third party can use the collected information to profile users and infer their interests, demographics, and lifestyle for a variety of purposes from fraud detection to targeted advertising. This is especially concerning if the tracked information reveals visits of websites about sensitive topics that could incur social stigma or legal repercussions, including information about certain diseases, services offered to undocumented immigrants, or abortion providers [80, 125, 142, 235]. The ability of third parties to collect personal data about website visitors explains why third-party requests are one metric of interest in web tracking measurement studies.

Data sharing among third parties

The problem is exacerbated by the common practice of third parties loading other third parties and sharing gathered user data with them via mechanisms including *cookie syncing* [63, 210, 280], which website owners and developers may not be aware of [235]. This practice is widespread with services that are available free of charge and include online advertising networks for monetization, such as the third-party comment tool Disqus [153]. The GDPR was found to have limited effect on these practices, most notably leading to a consolidation towards a smaller number of actors engaged in data sharing and, thus, increased centralization that might be even more harmful to website visitors' privacy [282].

Specific privacy risks

More specific privacy risks involved with third-party use, as well as available mitigations, differ by the type of functionality integrated via the third-party service [162]. In Section 2.1.5, we investigate common use cases for third-party services in web development, along with their specific privacy risks and mitigations.

2.1.4 *Mitigations*

General mitigations

Privacy risks associated with the use of third-party services can be generally mitigated on two levels: the *selection* of how to integrate the desired functionality and, when a third-party service is chosen, the *configuration* of the selected third-party service. Some mitigations might be available for all types of third-party content, while others are specific to certain types of functionality. We first describe general mitigations before delving into specific use cases, their associated privacy risks, and alternatives in Section 2.1.5.

Selection of a less privacy-invasive integration

On the selection level, a privacy-friendly choice might be to self-implement the desired functionality. This may be relatively easy to do in the case of, for example, a button that leads visitors to the website's social media presence or self-hosting JavaScript libraries or design resources, but a daunting task for more complex functionalities such as web analytics [162]. Middle ground can be achieved by using third-party software to be installed on the website's host system. Another option is to select a third-party service that promises to collect little data from website visitors; however, these alternatives may still transmit some data to the software vendor. As shown in Section 2.1.5, there are also some types of website functionality that require involvement of a third party, either by definition (e. g., social media integration) or for the sake of practicality (e. g., online payment).

Privacy-enhancing configuration

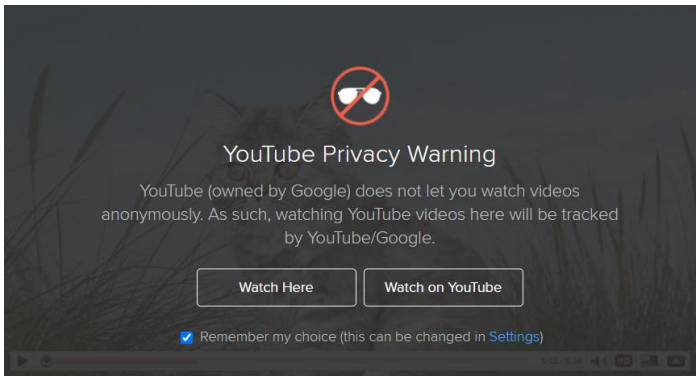
If a third-party service is selected, developers can potentially reduce the amount of personal data sent to the third-party server by selecting privacy-friendlier *configuration* options, if the service offers them. Examples include Google Analytics' feature for IP address anonymization[1] [101] or YouTube's "privacy-enhanced mode" [100] that prevents YouTube cookies to be automatically placed in visitors' browsers when they load the page with the embedded video, but only when they actually click to play it.

For some functionality, workarounds offered by parties other than the third-party vendor may be available, as outlined shortly in Section 2.1.5. One example, shown in Figure 2.1, are *two-click mechanisms* for embedded content, including videos or social media content. In these cases, data collection through the third party, such as the setting of cookies, is often already triggered when the page with the embedded content is loaded, not only when the visitor starts to interact with the embedded element. Two-click mechanisms aim to withhold this data collection: The original third-party content, which usually would have been automatically retrieved and displayed upon page load, is replaced with a placeholder that informs visitors about the content and the privacy implications of its inclusion and asks for their consent to proceed. There are also tools that

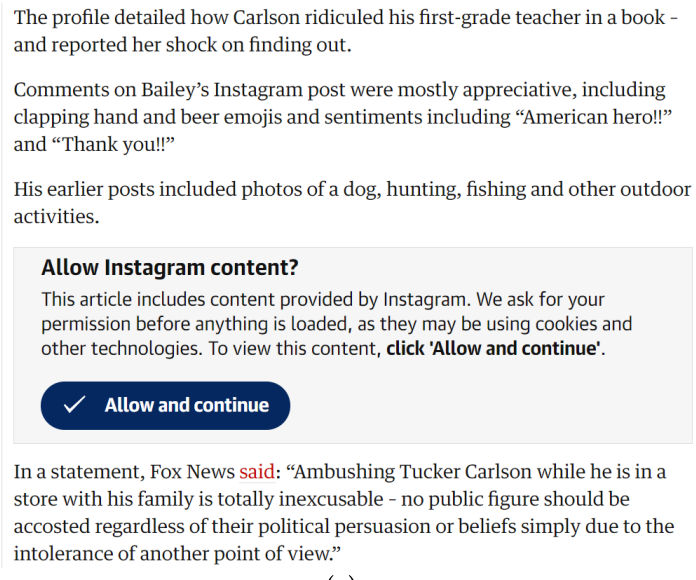
1 In the future, this will no longer be neccessary, as Google has announced to discontinue older versions of Google Analytics and shift users towards its version 4 that does not collect IP addresses [112].

Figure 2.1: Examples of two-click mechanisms to integrate third-party services such that data transmission to the third party is only triggered upon explicit user action: (a) two-click implementation of the Facebook "Like" button by Social Share Privacy [209], (b) search engine DuckDuckGo's preview of found YouTube videos, and (c) a consent button for an Instagram element embedded in the context of an article by The Guardian [116].

(a)



(b)



(c)

allow for the download and local hosting of third-party resources contrary to the third-party vendor's intentions, such as Google Fonts downloaders [73].

2.1.5 *Common Use Cases for Third-Party Services*

Comparing existing categorizations

Specific privacy risks of third-party use on websites depend on the type of functionality they provide, so it is first necessary to identify common use cases for third-party services in web development. We identified these through review and comparison of existing categorizations in the literature and by web tracking projects. We found such classifications in the works of Sørensen and Kosta [251], Libert and Nielsen [162], by WhoTracks.me [139], Third Party Web [121, 122], and DuckDuckGo's Tracker Radar [56].

While these categorizations differ in granularity and focus, we identified large overlap from the perspective of website owners. Table 2.1 shows the result of the comparison based on category names and definitions, along with our proposed consolidated version.

In the following, we describe the categories of third-party functionality from this consolidated list that will be relevant for this thesis, specifically Chapters 4 and 5. For each category, we provide a brief definition, explain specific privacy risks, and mention possible alternatives that do not collect as much personal information from website visitors.

ADVERTISING for third-party services or goods to generate revenue for the website. Online advertising is often implemented in the form of Online Behavioral Advertising (OBA), in which advertising spots on the publisher's website are auctioned off to interested third-party advertisers via Real-Time Bidding (RTB) based on the website visitor's presumed demographics and interests, collected and inferred via analysis of browsing habits, tracking techniques, and data sharing with other parties in the advertising ecosystem. Privacy-friendlier alternatives that do not rely on user data include contextual ads (e. g., EthicalAds [65]), static ads, or affiliate or sponsored content.

ANALYTICS measure visitors' behavior to evaluate website performance and marketing success. Collected data may include sites visited, links or areas clicked, bounce rates, or even screen or session recordings. The market leader, Google Analytics, is the most widely used third-party service on the Web [63, 121, 140]. It collects extensive data, might share it with Google's advertising network, and can track people's browsing behavior across websites. The service can be configured to use less data [98], e. g., using IP anonymization as required in some jurisdictions to be compliant with privacy law [145]. Alternatives include services that can be self-hosted or collect as little data as possible (e. g., Matomo).

EMBEDDED MEDIA refers to content such as videos, audio files, interactive maps, charts, or slideshows, embedded into web pages. Often the content is hosted by a third party (e. g., YouTube, Google Maps). Respective services typically provide code to embed a resource into a website. However, this code may result in visitor data being collected by the third-party

Table 2.1: Classifications of third-party services by purpose from related work and web tracking databases.

Consolidated	*WhoTracks.me* [91, 139]	*Sørensen / Kosta* [251]	*Hulce* [122]	*Libert / Nielsen* [162]	*DuckDuckGo* [56]
Advertising	Advertising, Adult Advertising / "Pornvertising"	Advertising	Ad, Marketing	Advertising and marketing, Content recommendation	Advertising, Ad Motivated Tracking
Embedded Media	Audio / video player	Content	Video, Content	Content hosting (video), Content hosting	Embedded Content
Content Delivery	CDN	Distribution technology	CDN	–	CDN
Hosting	Hosting	–	Hosting	Content hosting	–
Customer Interaction	Customer Interaction, Comments	–	Customer success	–	Social – Comment
Website Protection	–	Cybersecurity	–	–	Ad Fraud
Tag Management	Essential	–	Tag manager	–	Third-party Analytics Marketing
Login / Authentication	–	–	–	–	Federated Login, SSO
Payment	–	–	–	–	Online Payment
Privacy	Essential	Privacy	–	–	–
Programming / Design	–	Programming	–	Content hosting, Design optimization	–
Analytics	Site analytics	Analytics	Analytics	Audience measurement	Analytics, Audience Measurement, Third-Party Analytics Marketing, Action Pixels, Session Replay
Social Media	Social media	–	Social	Social media	Social Network, Social – Comment, Social – Share
Not mappable	Extensions	Editorial, Publisher, Retail, Plug-in	Utility	–	Badge, Malware
Other	Unknown, Misc	Unidentifiable	Other	–	Obscure Ownership, Unknown High-Risk Behavior, Non-tracking

service as soon as the visitor accesses a web page with embedded content. Some services offer privacy-friendly configuration to only transmit data if the visitor starts interacting with the embedded element [73]. Alternatively, there exist two-click solutions [206] created by entities other than the third-party vendor that only trigger embedded media to be loaded after consent. Media content can also be self-hosted or services can be used that do not store visitor data.

CUSTOMER INTERACTION functionality enables specific interactions between the website and its visitors. Examples are contact forms, comment sections, mailing lists, or chat boxes. These elements can contain a variety of personal data. Simple forms could be self-implemented or made available by a Content Management System (CMS), but third-party services are popular and include Google Forms, Facebook Comments, or Disqus. The latter has been found to share personal data of non-EU users with ad networks by default without notice [105]. Leakage of personal data to third parties has also been reported for contact form solutions [256].

USER LOGIN / AUTHENTICATION functionality allows websites to offer user accounts for visitors. It may be provided by a CMS, self-implemented, or made available via use of Single Sign-On (SSO) mechanisms. The latter allow the use of credentials from identity providers, such as Facebook, Twitter, Google, Apple, or GitHub, rather than having visitors create new credentials. The drawback is that these these providers learn on which websites people use their credentials and when [141].

PAYMENT functionality allows website visitors to pay for goods and services offered on the website. While methods exist that do not involve third parties (e. g., gift cards, cash) they are not particularly practical or widespread in an online context. Practical online payment solutions typically require both parties' banking institutions or an intermediary to be involved. While direct bank transfers are possible, third-party payment services are widely used, such as PayPal, Venmo, or Alipay, with the popularity of services differing across cultural regions [27]. This involvement of intermediaries comes with the sharing of customers' sensitive personal and financial data [219]. There is also the possibility that third-party payment services share data with other third parties not involved in the transaction [212].

PRIVACY NOTICES AND FORMS help fulfill transparency, consent, and opt-out requirements mandated by privacy legislation (e. g., ePrivacy Directive and GDPR in Europe; CCPA in the US). While it is possible to self-implement the necessary mechanisms, many websites use third-party services which promise compliance but have been found to not always be correctly implemented, allowing data processing to take place without valid consent [174], and frequently employ dark patterns to nudge website visitors to give consent [195]. In Chapter 3 we investigate the technical capabilities of selected third-party plugins for cookie consent.

PROGRAMMING AND DESIGN RESOURCES such as web fonts and front-end programming frameworks (e.g., jQuery, Bootstrap) are often directly

loaded from a third-party server. For widely used resources, there is the generic privacy risk of the provider potentially learning which other websites people visit that also include these resources. Including the resource from a central repository has the advantage that it may already be cached in visitors' browsers but has also been found to negatively impact website performance [216, 225] and could require users to accept comprehensive privacy policies, for example, Google's in the case of Google Fonts [73]. Alternatively, such resources could be self-hosted [73, 150, 216, 225].

SOCIAL MEDIA INTEGRATION connects a website with social media services, for example, via links to a website's social media profiles, buttons to share website content on social media, or embedded social media feeds. Similarly to embedded media, embed code provided by the social media service triggers data collection as soon as the page with the embed mechanism is loaded, but unlike that functionality this by definition requires a third party to be involved, the social media service. The European Court of Justice (ECJ) ruled that site owners can be held liable for the processing of the data social media companies collect through buttons or widgets on the first-party website [17]. Alternatives include simple (image) links or, for embedded content, the aforementioned two-click-solutions such as Social Share Privacy [209] shown in Figure 2.1 (a), Shariff [205], or Embetty [206], which only trigger data transmission to the social network if the visitor has clicked to activate the embedded element.

WEBSITE PROTECTION mechanisms aim to protect against (distributed) denial-of-service attacks, spam, or data scraping. Common mechanisms include CAPTCHAs, bot detection mechanisms based on text or behavioral analysis, or services like Cloudflare that act as a security proxy between the website and the visitor. The most prevalent anti-bot mechanism is Google's reCAPTCHA [24], which uses a wide range of behavioral data to distinguish humans from bots [51, 73, 201], in its Version 3 even without visual indication [241], and, thus, requires visitors to implicitly accept Google's extensive privacy policy [51]. Simpler and less intrusive mechanisms such as honeypots or easy mathematical problems are considered sufficient to protect against non-targeted spam, the majority of spam on the Web [51].

2.2 LEGAL BACKGROUND

With third-party services and their privacy implications now established, this section moves on to provide the background information about applicable privacy laws necessary for this thesis. It introduces the titular General Data Protection Regulation (GDPR) and key provisions we expected to have an influence on how websites use third-party services. Since the GDPR by itself did not yet achieve its goal of fully harmonizing EU data protection laws, we also introduce the ePrivacy Directive that governs the setting of cookies in website visitors' browsers, including those originating from a third party. On the other side of the Atlantic, a recent new piece of legislation, the California Consumer Privacy Act (CCPA), introduced a requirement for websites to let visitors opt out of the

sale of their personal information, which can be implemented through use of third-party services and also potentially affect third parties' data collection through websites.

2.2.1 *General Data Protection Regulation (GDPR)*

EU directives and regulations

In 2012, the EU started to take regulatory action to harmonize data protection laws across its member states. Existing data protection legislation comprised the Data Protection Directive (95/46/EC) [75] and the ePrivacy Directive (2002/58/EC) [76]. In contrast to EU *regulations*, whose provisions directly apply in each member state, EU *directives* are only binding as to the result, leaving the member states to decide upon the form and method to achieve this goal. Hence, the aforementioned directives are not directly applicable in each member state but require implementation of their requirements into national laws.

These national implementations differed widely, resulting in a complex landscape of privacy laws across Europe, as pointed out by Recital 9 of the GDPR. Some member states embraced stricter privacy laws and enforcement while others opted for lighter regulation. The General Data Protection Regulation (EU 2016/679) [78] was intended to overcome this situation and harmonize privacy laws throughout the EU, setting high and consistent standards for the processing of the personal data of the people within its jurisdiction. The regulation was proposed in January 2012, adopted on May 24, 2016, and its provisions became enforceable on May 25, 2018.

The GDPR introduced new legal obligations for entities that process personal information. As the HTTP protocol involves the transmission of information including IP addresses that are considered personal data in the EU (see Sections 2.1.3 and 2.2.1.1) the GDPR applies to the operators of web services and, therefore, was expected to impact the technical design of websites, what data they collect, and how they inform users about their practices. In the following we present selected key provisions of the GDPR that govern the use of third-party services by websites. A more detailed discussion of the regulation can be found in the legal literature [229]. Santos et al. conducted an in-depth analysis of legal sources and identified concrete legal-technical requirements for valid consent to data processing under European privacy law [236].

2.2.1.1 *Applicability*

Material scope

First it is important to clarify to what types of data processing the GDPR applies. Its material scope is laid out in Article 2:

> *"... the processing of personal data wholly or partly by automated means ...[,]"*

with "personal data" meaning (Article 4(1)):

> *"any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data,*

> *an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person[.]"*

Regarding online identifiers, EU institutions have clarified that IP addresses are considered personal data under the GDPR [67] as they allow for the, if only temporary, unique identification of the individual. Thus, the collection of a website visitor's IP address through a third-party plugin (*"automated means"*) as described in Section 2.1.3 falls under the material scope of the GDPR.

Territorial scope

Article 3 proceeds to define the regulation's territorial scope. Its first paragraph lays down that the GDPR applies to data processing activity of entities with an establishment in the EU, regardless of whether the processing takes place in the EU or not. The second paragraph extends the GDPR's territorial scope to data processors outside the EU if they process

> *"personal data of data subjects who are in the [European] Union [...], where the processing activities are related to:*
>
> *(a) the offering of goods or services, irrespective of whether a payment of the data subject is required, to such data subjects in the [European] Union; or*
>
> *(b) the monitoring of their behaviour as far as their behaviour takes place within the Union."*

Thus, from a territorial perspective, the GDPR governs any processing of personal data either by entities with a representation in the EU or for services offered in the EU, even if the service provider does not have any legal representation there. For online services this means that any website offering its services in the EU has to comply with GDPR standards. Due to this broad territorial scope, the GDPR affects millions of web services from around the world, as long as they are available in Europe.

2.2.1.2 *Transparency*

Among the core principles of the GDPR are its transparency requirements. Article 12(1) GDPR requires that anyone who processes personal data should inform the data subject, i. e., the person whose personal data is collected (Article 4(1) GDPR), about the processing and present the information

> *"in a concise, transparent, intelligible, and easily accessible form, using clear and plain language[.]"*

The typical means for this are privacy policies, which may be provided in a separate document, as parts of comprehensive terms of service, or even split into multiple documents, such as a designated cookie policy. Article 13 more specifically lists what information needs to be provided. This includes contact data, the purposes and legal basis for the processing, and the data subject's rights regarding their personal data, for example, the right to access, rectification, or deletion. These requirements make it necessary for every web service that offers its services to people in the EU to have a privacy policy and modify existing privacy policies to comply with the new transparency requirements.

2.2.1.3 *Legal Bases for the Processing of Personal Data*

Another core regulation of the GDPR is Article 6, which introduces a finite set of six legal bases. This means that the processing of personal data is only lawful

> *"[...] if and to the extent that at least one of the following applies:*
>
> *(a) the data subject has given consent to the processing of his or her personal data for one or more specific purposes;*
>
> *(b) processing is necessary for the performance of a contract to which the data subject is party [...];*
>
> *(c) processing is necessary for compliance with a legal obligation to which the controller is subject;*
>
> *[...]*
>
> *(f) processing is necessary for the purposes of the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of personal data [...]."*

Stricter conditions for lawful collection apply to sensitive categories of personal information including data about an individual's health, ethnic origin, or sexual preference (Article 9 GDPR).

Consent to the processing of personal data

Most important for this work is *consent* as the legal basis for data collection, which is further defined in Article 4(11) GDPR as

> *"any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her[.]"*

Conditions for consent

More concrete guidance on the four core requirements – freely given, specific, informed, and unambiguous – is provided in "opinions" issued by the European Data Protection Board (EDPB) [71] and its predecessor, the Article 29 Data Protection Working Party (A29DPWP) [14]. According to the opinion on consent under the GDPR [71], *freely given* means the data subject needs to be offered "real choice and control"; if they feel compelled to agree to the processing of their personal data or are threatened with negative consequences, this does not constitute valid consent. *Specific* means that consent under Article 6(1)(a) GDPR be given "for a specific processing purpose," which requires data subjects to be "specifically informed about the intended purposes of data use concerning them." As a consequence, web operators cannot collect valid consent to data collection for undisclosed or vague purposes. Though not explicitly mentioned in any provision, EU privacy law also demands that consent must be given *prior* to the data processing [236].

Subsequent GDPR articles contain additional requirements that go beyond the definition in Article 4(11) but can be understood to be conditions for the validity of consent [236]. These include the requirement for the processing

party (the "data controller") to be able to prove consent (Article 7(1)) and the data subject's right to withdraw consent at any time (Article 7(3)). For children under the age of 16 consent can only be given by the holder of parental responsibility (Article 8).

IAB Europe Transparency and Consent Framework

In preparing for the GDPR, the online advertising industry, which heavily depends on collecting and sharing Web users' personal data for the purpose of Online Behavioral Advertising (OBA), proposed its own solution to obtain consent to data processing from website visitors, the Internet Advertising Bureau (IAB) Europe Transparency and Consent Framework (TCF) [132]. Building upon the concept of consent notices that emerged in the wake of the ePrivacy Directive, as we will see shortly in Section 2.2.2, this framework aims to standardize how consent information is presented to the user, collected, and passed down the online advertising supply chain [132]. TCF-supporting Consent Management Platforms (CMPs) may display a list of third-party vendors participating in the framework, and the user can select which vendor should be allowed to use their personal data for a variety of different purposes. The user selection is encoded in a *consent string* and transmitted to the participating third-party vendors who committed to comply with the user's selection. Over the course of the research projects that serve as the basis for this thesis, the TCF found widespread use by websites [114], though research soon found its consent mechanisms often not to be correctly implemented [174]. In early 2022, the Belgian DPA issued a fine of €250,000 against IAB Europe on the ground that the TCF violated multiple GDPR provisions, including specificity of the provided information and accountability for obtained consent [22].

Legitimate interest to use third parties?

Beyond consent, an important question under the GDPR is to what extent the use of third-party services on websites can be based on the website operator's legitimate interest (Article 6(1)(f)) or whether visitors' consent is required (Article 6(1)(a)). Definite answers are only available if given by binding sources, such as EU courts; and, to a lesser degree of certainty, if non-binding sources such as the EDPB or national DPAs have issued assessments, which, in the case of the latter, may differ between member states. In January 2022 a mid-level court in Germany ruled that use of Google Fonts, the most popular remote font hosting service, cannot be based on legitimate interest (Article 6(1)(f) GDPR) on the grounds that it is possible to self-host the fonts [150]. Other concrete third-party services have been deemed not compliant with the GDPR due to transfer of personal data to third countries, as described in the following.

2.2.1.4 *Transfer of Personal Data to Third Countries*

The use of third-party functionality hosted on servers different from the first-party website may quickly lead to visitors' personal data being transferred across EU borders. Articles 44–49 GDPR define principles for the transfer of personal data to entities outside EU jurisdiction, including the general principle in Article 44 that requires that for any such transfer it needs to be ensured that

> *"the level of protection of natural persons guaranteed by [the GDPR] is not undermined."*

This is of particular interest if a website uses third-party services whose servers are located in the United States, as previous agreements between the EU and the US regarding adequate protection of EU citizens' personal data had been declared void by the ECJ due to insufficient protection against US mass surveillance programs [68, 70]. This argument served as the basis for two decisions by the Austrian and French DPAs in early 2022 that declared use of the most prevalent third-party service on the Web, Google Analytics [198, 217], not compliant with the GDPR. In late March 2022 a new EU–US data protection agreement was announced to be in the works [177].

2.2.1.5 *Data Protection by Design and by Default*

Article 25 GDPR contains important core principles of the regulation. It states that entities processing personal data should,

> *"taking into account [...] the state of the art, the cost of implementation and the nature, scope, context and purposes of processing [...,] implement appropriate technical and organisational measures, such as pseudonymisation, which are designed to implement data-protection principles, such as data minimisation, in an effective manner [...]."*

The "state of the art" in data protection

The "state of the art" by definition is subject to change. European DPAs and industry associations in information security have issued tech reports that attempt to describe this "state of the art." One explicitly named web security mechanism is transport encryption of HTTP via Transport Layer Security (TLS), i. e., HTTPS [130, 270]. Over the last couple of years its adoption has increased, fueled by more readily available certificates and major browsers increasingly flagging HTTP-only connections as insecure [10], so that the majority of today's Web traffic is transport encrypted (as of 2017/18, between 59 and 89 % depending on traffic analysis method; see Felt et al. [82] and our own findings in Chapter 3).

Thus, use of HTTPS can be considered mandatory under the GDPR, particularly for web pages that directly prompt for personal information, such as contact forms or payment services. A similar reference to the "state of the art" can be found in Article 32 GDPR ("Security of processing").

Further, Article 25(2) GDPR requires data controllers to

> *"implement appropriate technical and organisational measures for ensuring that, by default, only personal data which are necessary for each specific purpose of the processing are processed. That obligation applies to the amount of personal data collected, the extent of their processing, the period of their storage and their accessibility."*

In the context of third-party services on websites this can be interpreted as a legal obligation for websites to integrate a desired functionality in such a way that only the data necessary for that functionality is collected. This often is not the case, as shown in Section 2.1.4 through the existence of alternatives to popular third-party services that collect less user data.

2.2.1.6 *Fines*

One factor that contributed to the broad discussion of the GDPR's impact on the Web economy is the possibility for supervisory authorities to impose significant fines, as laid out in Article 83(5):

> *"Infringements of the following provisions shall ... be subject to administrative fines up to 20,000,000 EUR, or in the case of an undertaking, up to 4 % of the total worldwide annual turnover of the preceding financial year, whichever is higher:"*

The text then proceeds to list the GDPR principles for whose violations such fines can be imposed. Of particular interest to this thesis are the following cases:

> *"(a) the basic principles for processing, including conditions for consent, pursuant to Articles 5, 6, 7 and 9";*
>
> [...]
>
> *"(c) the transfers of personal data to a recipient in a third country or an international organisation pursuant to Articles 44 to 49"; [...]*

This refers to the aforementioned requirements regarding a legal basis for the data processing, including consent, and data transfers across EU borders. In Chapter 6 we investigate if the prospect of fines can motivate web operators to fix faulty implementations of consent on their websites.

2.2.2 *ePrivacy Directive*

In an earlier effort to update EU privacy laws for the digital age, Directive 2009/136/EC [77] applied changes to the ePrivacy Directive (2002/58/EC) [76], the existing piece of legislation created on the EU level to address data protection and privacy in electronic communication systems including the Internet. The changes include the introduction of a consent requirement to store HTTP cookies in users' browsers. Due to this prominent change, the updated directive is often colloquially referred to as the "(EU) Cookie Directive."

Consent to the use of cookies

In its 2009 version, Article 5(3) of the ePrivacy Directive states tasks member states to

> *"ensure that the storing of information, or the gaining of access to information already stored, in the terminal equipment of a [...] user is only allowed [if] the [...] user [...] has given his or her consent, having been provided with clear and comprehensive information [...] about the purposes of the processing."*

Exception for "strictly necessary" cookies

This rule is followed by an exception: The consent requirement for the storage or access of cookies in website visitors' browsers does not apply if the stored information is used

> *"for the sole purpose of carrying out the transmission of a communication over an electronic communications network, or as strictly necessary in order for the provider of [a] service explicitly requested by the [...] user to provide the service."*

Like for consent under the GDPR, the A29DPWP has provided guidance on consent under Article 5(3) of the ePrivacy Directive [12], as well as exemptions [11]. Consent under the ePrivacy Directive is subject to similar requirements as under the GDPR; it requires a freely given, active choice based on specific information about the purpose of the processing and given before the processing starts [12]. In the exemption clause, the second case is more relevant for websites; it is understood to exempt cookies from consent if an explicitly requested service "with a clearly defined perimeter" would not work without setting the cookie [11]. Examples include cookies remembering the state of the shopping cart in an online shop or the fact that the user has logged into their account. The ePrivacy Directive's rules for cookie consent can apply to the use of third-party services on websites, as many of them set cookies to remember visitors and their preferences across websites. As the purpose of third-party cookies tends to be distinct from the explicitly requested service on the first-party website and they are often persistent to facilitate tracking, the A29DPWP opinion found third-party cookies to be unlikely to fall within the bounds of the exemption.

Implementation deficits

Directive 2009/136/EC became effective in 2011 and required each member state to implement its provisions into national law by May 25, 2011. A 2015 study by London-based law firm Bristows found that even four years later, the status of implementation across EU member states, as well as the interpretation of Article 5(3), greatly varied [20].

(Cookie) consent notices

Still, the effects of this legislation soon started to show on European websites in the form of *(cookie) consent notices*, often referred to as *cookie banners* – popup boxes or banners shown to new visitors of a website informing them about the use of cookies and similar tracking technologies by the website and associated third parties. Consent notices emerged from practice and were later acknowledged by the A29DPWP as a possible component of a mechanism for cookie consent [12], though the widespread practice of implying consent from continued use of the website fell short of the A29DPWP's requirements for valid consent under Article 5(3). It has to be noted that Article 5(3) applies to any kind of information stored on the user's system, even if it does not contain any personal information.

2.2.3 *Planned ePrivacy Regulation*

No updated rules on cookie consent

The GDPR deliberately did not address consent to the use of cookies, leaving the question for the planned *ePrivacy Regulation*, another EU regulation designed to complement the GDPR and complete the harmonization process. It was originally supposed to become effective at the same time, in May 2018. Due to ongoing disagreement between EU institutions and member states about its regulatory depth, the legislative process for the ePrivacy Regulation was delayed multiple times. As of July 2022, its future is still unclear, but the latest proposal from January 2017 [66] contains additional exemptions from cookie consent in its Article 8, including for first-party analytics.

GDPR vs. ePrivacy Directive

For the time being, the EDPB has clarified the relationship between the ePrivacy Directive and the GDPR for the use of cookies, as follows: Article 5(3) of the ePrivacy Directive governs access to non-necessary cookies in the user's browser, whether they contain personal data or not, while the GDPR applies to subsequent processing of personal data retrieved via cookies [72].

2.2.4 *California Consumer Privacy Act (CCPA)*

Recent years have also seen countries outside the EU update their privacy laws. One prominent example with ramifications for websites is the California Consumer Privacy Act (CCPA) [258], incorporated in Section 1798.100 of the California Civil Code (CCC). The CCPA's goal is to strengthen the privacy rights of consumers in the US state of California. It was passed on June 28, 2018 and became effective on January 1, 2020.

Right to opt out of the sale of personal information

Among the granted rights are certain rights of consumers to limit the sale of their personal information through a business to third parties, as constituted in Section 1798.120 CCC:

> *"A consumer shall have the right, at any time, to direct a business that sells personal information about the consumer to third parties not to sell the consumer's personal information. This right may be referred to as the right to opt out."*

Section 1798.135 CCC provides more detail how consumers need to be made aware of this right:

> *"(a) A business that is required to comply with Section 1798.120 shall, in a form that is reasonably accessible to consumers:*
>
> *(1) Provide a clear and conspicuous link on the business' Internet homepage, titled "Do Not Sell My Personal Information," to an Internet Web page that enables a consumer [...] to opt out of the sale of the consumer's personal information. A business shall not require a consumer to create an account in order to direct the business not to sell the consumer's personal information.*
>
> *(2) Include a description of a consumer's rights pursuant to Section 1798.120, along with a separate link to the "Do Not Sell My Personal Information" Internet Web page in:*
>
> *(A) Its online privacy policy or policies if the business has an online privacy policy or policies.*
>
> *(B) Any California-specific description of consumers' privacy rights."*

Definition of "sell"

It has to be noted that "sell" does not necessarily involve the transfer of monetary funds in exchange for the personal information, as defined in Section 1798.140(t)(1):

> *"'Sell,' 'selling,' 'sale,' or 'sold,' means selling, renting, releasing, disclosing, disseminating, making available, transferring, or otherwise communicating orally, in writing, or by electronic or other means, a consumer's personal information by the business to another business or a third party for monetary or other valuable consideration."*

The definition of "consideration" in California law allows for the interpretation that any benefit a business receives through the disclosure of personal information constitutes "valuable consideration" and, thus, a sale. Based on this, it has been argued that the use of cookies for third-party analytics and advertising provides businesses with sufficient benefits to constitute a "sale" and trigger the opt-out requirement [16].

"Do Not Sell" link

In the wake of this legislation, websites targeted at Californians have begun to include the required "Do Not Sell My Personal Information" link that usually leads to a contact form where consumers can provide their details and request the opt-out. This is a rare example where privacy legislation provides concrete instructions how to implement one aspect of a certain legal requirement. Still, first research results show that despite this, the actual wordings websites use for the link greatly vary [199].

IAB CCPA Compliance Framework

The widespread need for such functionality has prompted some vendors of third-party privacy plugins, such as consent notices, to also offer services that implement the "Do Not Sell" requirement. Similar to the IAB Europe TCF, the IAB's main (global) branch has created its IAB CCPA Compliance Framework [131] that specifies a "US Privacy String" that stores users' decisions regarding their opt-outs to the sale of their personal information and passes it to other actors in the online advertising ecosystem.

Part II

WEBSITE VISITORS' PERSPECTIVE: CONSENT NOTICES

3

PRIVACY-RELATED CHANGES ON WEBSITES AROUND THE GDPR ENFORCEMENT DATE

3.1 INTRODUCTION

GDPR principles and presumed effect on websites

In Section 2.2.1 we established core principles of the GDPR, including its mandate for "data protection by design and by default," extensive transparency requirements, and the need for a legal basis for data collection. As a result, its presumed impact on online businesses and both commercial and personal websites in the light of possible fines had been widely discussed in the months leading up to the GDPR enforcement date on May 25, 2018, prompting websites to notify users of new privacy policies or terms of service, restrict services offered to EU visitors, or (temporarily) cease operations in the EU altogether [113]. This raised the expectation that the regulation – and its associated fines – becoming enforceable would prompt websites' data processing practices to change, including decreasing use of third-party services that unnecessarily collect visitor data, greater transparency in the disclosure of data processing practices, and increased prevalence of state-of-the-art data protection technology.

Role of web privacy measurements

Measuring and monitoring such changes at scale can inform regulators about the effects of privacy legislation in practice and assist enforcement through detection of widespread implementation deficits. While data protection authorities have conducted such assessments in the past, for example, about the state of the implementation of cookie consent according to Article 5(3) ePrivacy Directive [13], they are often limited by available human and financial resources.

While web privacy research had already studied the evolution of web tracking over time [155], including use of third-party services [288], the GDPR coming into effect provided the unique opportunity to study and compare websites' privacy practices before and after significant changes to data protection legislation that were expected to directly affect these practices.

Previous work studying websites' disclosures about data processing had focused on privacy policies as the main mechanism through which businesses inform about their processing of personal information. A rich body of work has studied privacy policies on websites from a variety of different angles, including their prevalence (70 to 80 % of company websites in the US [194]), readability [179, 224], and, more recently, automated content analysis and extraction of data collection practices [110, 165, 271, 295], including disclosure of third-party use [161]. By contrast, cookie consent notices and associated cookie policies are a more recent phenomenon rooted in the ePrivacy Directive's 2009 revision, which member states were slow to implement into their respective national laws (see Section 2.2.2). Consequently, as of early 2018 consent notices had only seen previous research attention with respect to their usability [147] and first hints at their prevalence [13], but their use and implementations had not yet been studied in detail.

Our contribution: longitudinal measurements of privacy-related changes

In this chapter, we address this research gap and conduct an empirical study to measure changes that occurred on a representative set of websites around the time the GDPR came into force. We took the opportunity to study the impact of this rare event and monitored the 500 most popular websites in each of the 28 member states of the EU[1] over the course of eleven months. In total, this resulted in a set of 6,759 websites in 24 different languages. We used a combination of automated and manual methods and tracked these websites' transparency mechanisms (cookie consent notices and privacy policies), as well as selected tracking metrics (use of third-party services and cookies) and a data protection mechanism (use of HTTPS) before and after the GDPR enforcement date, all with the goal to identify changes that could possibly be attributed to the new legislation.

Our results show that changes made around the GDPR enforcement date had an overall positive effect on the transparency of websites' privacy practices: For web users in Europe, the most visible change was an increase in the prevalence of cookie consent notices. On average, 62.1 % of the analyzed websites used such notices after the GDPR enforcement date, while only 46.1 % had done so in January 2018. In order to better understand this phenomenon, we manually inspected 9,044 domains for their use of cookie consent notices and evaluated 28 common cookie consent libraries for properties useful to implement GDPR-compliant consent. We found that existing implementations greatly vary in functionality, especially the granularity of control offered to the user and the backend required to correctly implement the desired cookie configuration.

This increase in transparency is contrasted by limited change in websites' actual privacy practices: We could not find a significant reduction in tracking through third-party cookies, and the majority of sites relied on opt-out consent mechanisms, making the use of cookies and similar tracking technology the default for visitors. We found only 37 sites that asked for explicit consent before setting cookies. As an indicator for the use of state-of-the-art data protection technology we monitored websites' use of HTTPS. We found some slightly increased activity in Hypertext Transfer Protocol Secure (HTTPS) adoption around the GDPR adoption date, though this followed the general trend towards increased HTTPS use.

In summary, the work presented in this chapter makes the following contributions:

1. In order to study how websites react to the GDPR, we conduct an empirical, longitudinal study of websites' transparency mechanisms and selected data processing practices on a set of 6,759 websites, composed of the 500 most popular websites in each of the 28 EU member states. From January to October 2018, we performed monthly website crawls to measure changes in privacy-related metrics, including use of third-party cookies, cookie consent notices, and HTTPS adoption. Between January and the end of May, we observed an average increase in the prevalence of cookie consent notices by 16 percentage points.

1 This research was conducted in 2018, i. e., before the United Kingdom (UK) left the EU in 2020, thus reducing the number of its member states to 27.

2. We compare the use of cookies and third-party services in our set of websites between January and June 2018 to determine whether the GDPR's transparency and consent requirements affected the prevalence of web tracking. While both were not significantly impacted, 147 sites stopped using tracking libraries and 37 chose to ask for explicit consent before activating them. For HTTPS adoption, we observed a slight increase around the GDPR enforcement date that lay within the general trend towards use of HTTPS over HTTP.

3. We categorize cookie consent notices based on their options for user interaction and identify six categories. We measure their prevalence by country and find that only a minority provides website visitors with an actual choice to allow or deny the use of cookies and similar tracking technology. We investigate this observation in more detail by taking a closer look at the many distinct implementations of cookie consent notices we found in our data set. We analyze these libraries for key features required to implement the legal requirements for consent under the GDPR and ePrivacy Directive and identify technical obstacles to achieving this goal.

Analysis of privacy policies

Beyond consent notices, the work that serves as the basis for this chapter also prominently investigated privacy policies as websites' main transparency mechanism regarding the collection and use of their visitors' personal information. The methods and results of the privacy policy analysis will be included in Henry Hosseini's PhD thesis and can also be found in the source paper [52].

3.2 RELATED WORK

Existing work has taken a first look at cookie consent notices. In addition, there is research about related mechanisms that offer website visitors the option to control the personal information websites collect about them.

3.2.1 *Cookie consent notices*

Prior to the work that serves as the basis for this chapter, research on cookie consent notices was scarce.

In February 2015, the Article 29 Working Party conducted a "Cookie Sweep" to determine the effects of Directive 2009/136/EC's requirements [13]. In eight EU member states, 437 sites were manually inspected for information they provided about cookies, including the type and position of the interface used. At that time, 116 (26 %) of the analyzed sites did not provide any information about cookie use; for another 39 % the information was deemed not sufficiently visible. Of the remaining 404 sites, 50.5 % (204) sites were found to *"request [...] consent from the user to store cookies"* while 49.5 % (200) simply stated that cookies were being used. 16 % (49 sites) offered the user to accept or decline certain types of cookies. The study did not investigate whether the banners asking for consent implemented a proper opt-in mechanism.

In November 2017, Kulyk et al. [147] collected cookie consent notices from the top 50 German websites in the Alexa ranking to investigate how users perceive and react to different types of banners. They identified five distinct groups

of notices based on the amount of information they provide about cookie use but did not analyze users' options for interacting with the banner. Participant sentiments when encountering a consent notice varied from disturbance, privacy concerns, and annoyance due to a lack of options to habituation and a lack of information. The wordings of the notices did not have a significant impact on visitors' decision to stay on the page or leave it.

3.2.2 *Mechanisms to Control Data Processing on Websites*

In the past, different technical solutions have been proposed and studied to help users cope with the ever-growing number of online tracking and profiling services.

P3P

In 2002, the P3P Project [45] was officially recommended by the W3C. It relied on machine-readable privacy policies directly interpreted by the browser, which was enabled to automatically negotiate, for example, the handling of certain cookies based on the user's preferences. However, none of the major web browsers support P3P anymore due to a lack of adoption by the industry, and consequently its W3C working group was closed in 2006 [44].

Do Not Track

Another approach was the Do Not Track (DNT) Header for the HTTP protocol, proposed in 2009 [83]. DNT was supported by all major browsers and allowed the user to signal online content providers their preference towards tracking and behavioral advertising. However, many websites did not honor DNT signals, as evidenced by a 2015 study that found no significant difference between visiting websites with the DNT header and without any tracking protection [64]. Due to this lack of adoption, the W3C discontinued its Tracking Protection Working Group in 2019, and Apple removed DNT support from Safari in 2019, stating the header could be used as a feature for browser fingerprinting [249].

YourAdChoices

Companies in the Online Behavioral Advertising (OBA) business point Web users to self-regulation programs such as *YourAdChoices* [55] to opt out of targeted advertising. Website visitors are informed of this option via a small blue icon in the corner of a displayed ad and shown additional information on click. For users this remains challenging, as studies have shown that they can hardly distinguish between different OBA companies [154] and have problems to even recognize and locate the corresponding icons [90].

We extend this existing work by analyzing the effect of the GDPR on websites' transparency mechanisms, particularly cookie consent notices, and investigate to what degree they offer visitors the opportunity to control tracking through third-party services. We also monitor changes in websites' actual tracking practices as indicated by the use of third-party cookies and measure HTTPS adoption over time as a metric for the use of state-of-the-art data protection technology.

3.3 METHOD

Combination of automated and manual analysis

To analyze the impact of GDPR enforcement on websites in the EU, we used automated tools combined with manual verification and annotation of websites

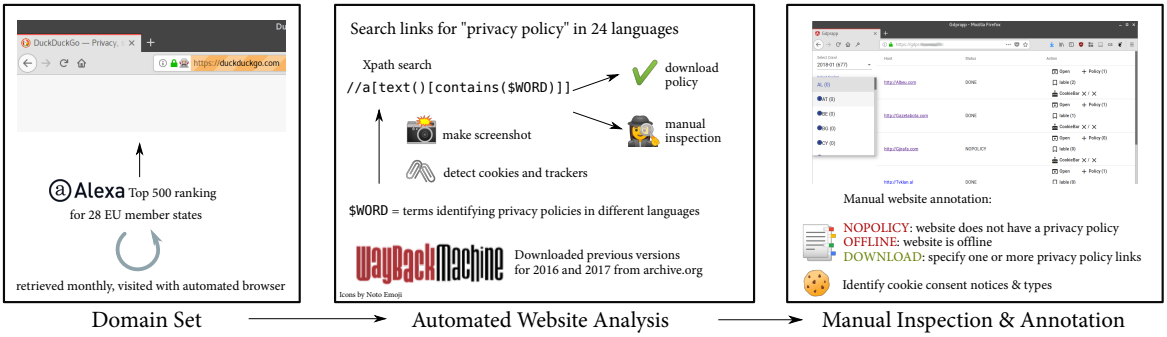


Figure 3.1: Overview of the website analysis process consisting of automated analysis, manual validation, and annotation.

in 24 different languages. We used an automated browser setup to periodically analyze websites for consent notices, privacy policies, and selected tracking metrics, plus a web app to allow for manual review and annotation of websites for consent notices, privacy policies, and thematic categorization. Figure 3.1 provides an overview of the main components of our website analysis system.

3.3.1 *Data Set Creation*

To create a diverse data set of popular websites in the EU, we leveraged the rankings of the 500 most popular websites by country as previously provided by Amazon's subsidiary Alexa[2] [6]. This service was selected because it was the only domain ranking service we were aware of that provided rankings by country at that time.

We started with a domain set that contained the top 500 websites for each of the 28 EU member states as of December 2017. To extend the scope of our study, we retrieved updated website rankings once per month and included newly added domains in our data set, thus continuously increasing the size of the domain set to crawl each month.

3.3.2 *Automated Website Checks*

Base setup and tracking detection

Our automated web browser was set up in a German data center with the Selenium WebDriver [250] using the latest available version of Firefox (version 57 onward) on servers running Ubuntu Linux and an XServer, so that all pages were actually rendered. Once a website's homepage had been loaded and completely rendered by the browser, we searched it for domain names of third-party advertising and tracking libraries based on EasyList [60], which is often used in popular ad-blocking browser extensions.

Use of HTTPS

To obtain another metric for whether websites over time adopted GDPR principles such as use of state-of-the-art techniques for data protection, we investigated websites' adoption of HTTPS by default. For this, our automated

2 Amazon retired Alexa Internet's services on May 1, 2022, but past URLs providing the domain rankings by country can be accessed via the Internet Archive's Wayback Machine [134], e. g., `https://web.archive.org/web/20220323041252/https://www.alexa.com/topsites/countries`.

browser would always try to resolve the HTTP address of a host and observe whether the visited website automatically redirected to HTTPS.

Finally, a screenshot of the rendered homepage was made to allow for manual inspection for cookie consent notices. As shown in Figure 3.1 and not further elaborated in this thesis, an important component of our automated website checks was the retrieval of websites' privacy policies.

Study timing

After a successful pretest of our crawling setup in December 2017, the websites were visited once per month from January to April 2018, three times in May (two times before and once after May 25, 2018) and again once per month until October 2018, resulting in 12 crawls in total. The results were stored in a MongoDB database.

3.3.3 *Manual Review*

In order to facilitate inspection for cookie consent notices (and validate the results of the automated detection of privacy policies), we implemented a web-based annotation tool to review and further process the collected data.

Manual inspection was performed with off-the-shelf browsers by four of the authors of the conference paper that serves as the basis for this chapter. Websites in languages unfamiliar to the annotator were translated with Google Translate. Automated translations through Google were available in all encountered languages and were good enough to figure out the general topic of a website and, paired with common design principles, whether it displayed a consent notice.

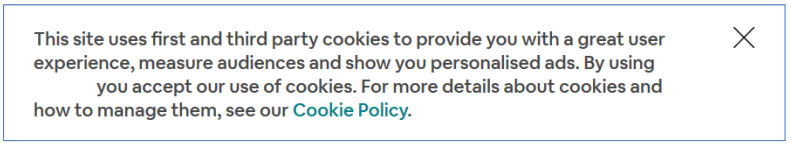
Websites that could not be accessed were labeled *Offline*. Under this label we merged all sites that were not reachable, occupied by a domain grabbing service, produced a screen indicating that the website was not available because of the detected location of our IP address, or belonged to a discontinued or not publicly accessible service.

3.3.4 *Categorizing Consent Notices*

In January, May, and again in early fall 2018, we manually inspected all websites for cookie consent notices. In January, we only noted whether a website displayed a consent notice or not. As we observed consent notices to become increasingly sophisticated over time, in the two other annotations we also categorized consent notices based the options for interaction they provided to website visitors. We identified the following distinct types, with examples shown in Figure 3.2:

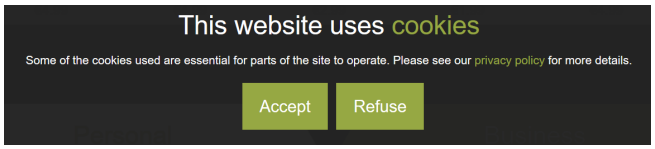
NO OPTION: Cookie consent notices with *no option* (Figure 3.2 (a)) simply inform visitors about the site's use of cookies. Visitors cannot explicitly consent to or deny cookie use. This category also includes banners that feature a clickable button whose label cannot be considered to express agreement (e. g., "Dismiss," "Close," or just an "X" to discard the banner).

CONFIRMATION: In contrast, *confirmation-only* banners (Figure 3.2 (b)) feature a button with an affirmative text such as "OK" or "I agree"/"I accept" that can be understood to express the visitor's consent.

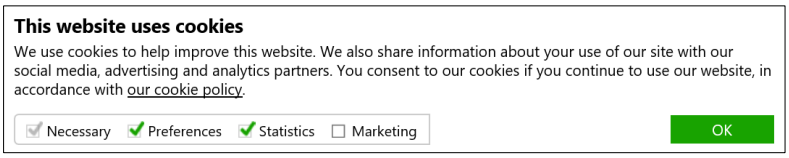


(a) No option

(b) Confirmation only

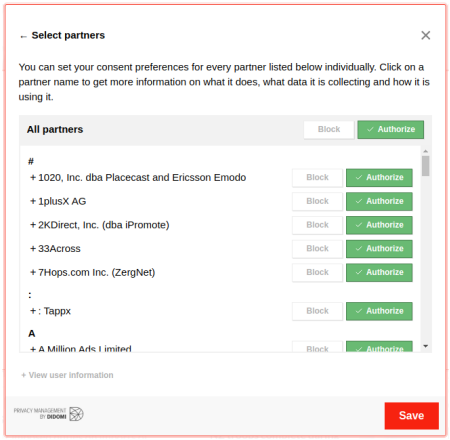


(c) Binary



(d) Categories

(e) Slider



(f) Vendors

Figure 3.2: Consent notices with different options for visitor interaction. First-party company names have been redacted.

BINARY consent notices (Figure 3.2 (c)) have two buttons that let visitors either explicitly agree to or decline all the website's cookies.

SLIDER: More fine-grained control is offered by consent notices that group the website's cookies into categories, mostly by purpose. *Slider*-based notices (Figure 3.2 (e)) arrange these categories into a hierarchy. The user can move a slider to select the level of cookie usage they are comfortable with, which implies consent with all the previously listed categories.

CATEGORY -based notices (Figure 3.2 (d)) allow users to accept or deny each category individually. The number of categories varies, ranging from 2 to 10 categories in our data set; we observed that most notices of the "checkbox" type featured 3–4 different cookie categories. A common set of categories comprises advertising cookies, website analytics, personalization, and what is usually referred to as (strictly) necessary cookies, such as shopping cart cookies. As described in Section 2.2.2, this type of cookies does not require explicit user consent under Article 5(3) of the ePrivacy Directive.

VENDOR: We assigned this category to banners that allow users to toggle the use of cookies for each third party individually. Figure 3.2 (f) shows one such mechanism. We first observed this type of selection on notices associated with IAB Europe's Transparency and Consent Framework (TCF) [132], which refers to third-party partners of the notice-using website as "vendors."

OTHER: This category, assigned five times in total, was used for consent notices that did not match any other category. For example, one site allowed users to choose between two "cookie profiles."

3.3.5 *Analysis of Cookie Consent Libraries*

During manual website annotation, we noticed company logos or "powered by" statements on consent notices, which hinted at the existence of third-party implementations to provide cookie consent notices. This raised questions about how common certain cookie consent solutions were and to what degree they could help website owners comply with the ePrivacy Directive and the GDPR.

Identifying consent libraries

We compiled a list of the cookie consent libraries identified during manual annotation. If possible, we downloaded each library or requested access to a (demo) account from the vendor. We implemented each consent solution – one at a time – into a live WordPress website. We then visited the site using Microsoft Edge 41 configured to not block any cookies, interacted with the cookie banner, and used Edge's Developer Console to observe the effect of user selection on the cookies stored to the machine. This analysis was conducted in August 2018, using each library's latest version available at that time.

Analyzing consent libraries

For each library, we tested the user interfaces it offered and whether its settings and documentation allowed us to block and unblock cookies (i. e., we did not write any custom code to implement new core functionality). We also tested if the libraries provided mechanisms to reconsider a previous consent decision and to log and store visitors' consent, as required by Article 7 GDPR.

It is in the interest of web service providers not to display consent notices to users that are not subject to the GDPR. Thus, many libraries offer the option to display the notice only to users accessing the site from specific regions of the world. We tested these location-dependent features using Tor Browser with an exit node in a country for which the consent notice was configured not to show up.

Popularity of consent libraries

We measured the popularity of identified cookie consent libraries in separate checks of domains' home pages in July and December 2018. To determine if a website used a cookie library, we reviewed the default locations of JavaScript and CSS resources and likely variants based on the installation instructions. Additionally, we checked for requests to third parties used by the libraries. We manually verified this procedure with a list compiled during the manual annotation phase. To reflect the exposure a library or service had to end users, we calculated a score based on the rank of the domain in Alexa's EU country rankings. This favors domains which are highly ranked in many country rankings over domains which only occur in a single country ranking. This better accounts for the exposure a library has to end users. This $Score$ inherits the general bias of the Alexa rankings (see Section 3.3.6). It is calculated by subtracting the $Rank_{country,i}$ of a domain from 501 for each country ranking (N) and summing up these values. Sites no longer present in the country rankings were assigned rank 501. The $Score$ is then normalized by dividing by N:

$$Score = \frac{\sum_{i=1}^{N} 501 - Rank_{country,i}}{N}$$

3.3.6 *Limitations*

Our measurement and analysis methods, as described above, have some limitations.

Bias in domain popularity rankings

Scheitle et al. [239] – and, concurrent to this study, also Le Pochat et al. [152] – showed that many publicly available website rankings, including Alexa, are biased, fluctuate highly, and that there are substantial differences among lists. Indeed, we observed high fluctuation, as countries' Alexa rankings from January and May only had, on average, 387 entries in common. Nevertheless, we relied on Alexa's website rankings, as they were the only available source for country-specific rankings. We accounted for high fluctuation by refraining from analyzing correlations between the websites' ranks and other measured factors, except for the impact of consent notice libraries. We accounted for bias potentially introduced through the domain lists by conducting the pre-post analysis only on domains present in the January rankings. To account for potential manipulation of rankings [152], especially in the light of some countries' small population, we excluded domains that were offline during one of the crawls or had been blocked by browser security mechanisms. Besides, the obligation to comply with legal regulations is independent of whether a site is legitimately listed in a top websites ranking or not.

Automated browser setup

The main data collection process (see Figure 3.1) was conducted with automated browsers using a server hosted in a known server farm. It is known that some websites change their behavior when an automated browser or specific server IP addresses are detected. We observed that several websites using

Cloudflare's services (see "Website protection" in Section 2.1.5) blocked direct requests and asked to solve a CAPTCHA before redirecting to the actual site. As described above, we checked for these effects when we manually visited all websites for annotation. Another drawback of our technical setup was that some websites might have changed their default language based on the server's IP address (Germany) or the default browser language (English). While this might have influenced the language of the presented consent notice and privacy policy, it should not have changed the fact that either existed.

3.4 RESULTS

Final domain set

In total, the Alexa rankings of the 500 most frequently visited websites for all 28 EU member states in January 2018 contained 6,759 different domains; the final list in November comprised 13,458 domains. Unless mentioned otherwise, pre-/post-GDPR comparisons are based on the data points for the domains first annotated in January, while the analysis of the cookie consent notices is based on the extended list we had created by the end of May.

Overview of findings

We did not find any significant change in websites' use of third-party services and cookies, as well as HTTPS adoption, before and after the GDPR enforcement date. The most notable (and visible) effect we observed is an increase in the use of cookie consent notices, from 46.1 % in January to 62.1 % in May. We found that especially popular websites use third-party libraries to implement cookie consent notices and the underlying backend. Our in-depth analysis of common libraries found in our dataset revealed shortcomings in how those consent mechanisms can satisfy the requirements of Article 6 GDPR and Article 5(3) of the ePrivacy Directive, as outlined in Sections 2.2.1.3 and 2.2.2.

3.4.1 *Adoption of "Data Protection by Default and by Design" Principles*

As a metric for how websites' actual data processing practices changed around the GDPR enforcement date, we investigated the presence of third-party tracking and HTTPS adoption over time.

3.4.1.1 *Tracking and Cookies*

Our automated website analysis did not reveal significant change in the prevalence of third-party tracking services or associated cookies. In January, websites used on average 3.5 third-party tracking services that would be blocked by an off-the-shelf ad blocker. In May, right before the GDPR came into effect, and in June we measured the number of first- and third-party cookies a website sets by default. Regarding third-party cookies no effect is visible; websites set about 5.4 cookies on average. The number of first-party cookies decreased from 22.2 to 17.9 cookies on average. This effect can be explained by a decrease in first-party cookie use in Croatia (-11.3) and Romania (-21.1). The medians stayed the same for both cookie groups.

Still, some websites made notable changes: We manually checked websites that did not use trackers in June but did so in January and found that 146 had

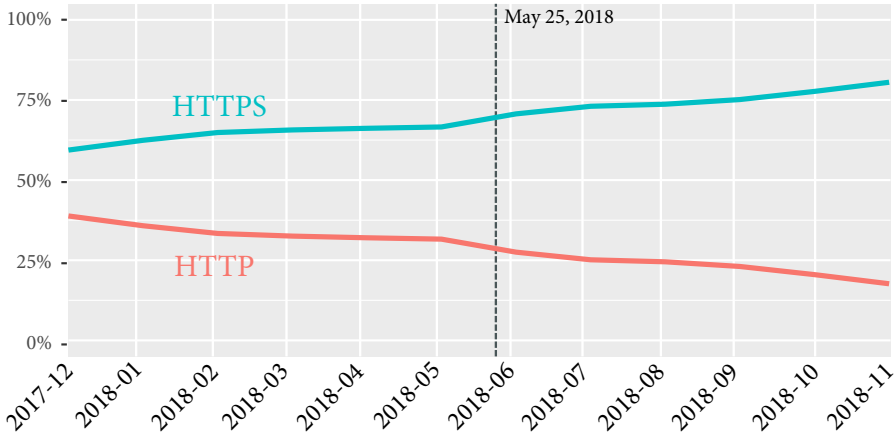


Figure 3.3: Changes in the adoption of HTTPS by default over time. The dotted line marks the GDPR enforcement date.

stopped using ad or tracking services altogether and 37 only tracked visitors after explicit consent. Notable examples included `washingtonpost.com` and `forbes.com`. Only after consenting to tracking – or subscribing to paid services – visitors were directed to the regular homepage of these sites.

3.4.1.2 *HTTPS Adoption*

As another metric for GDPR adoption by websites we measured their use of HTTPS by default, which can be considered "state of the art" in data protection mechanisms (see Section 2.2.1.5).

Increase within a larger trend

Our measurements confirm a general trend towards HTTPS adoption that had been reported before [82]. Figure 3.3 shows an increase in the use of HTTPS by default from 59.9 % in December 2017 to 80.2 % in November 2018. At the end of May, 70.8 % of websites redirected to HTTPS, close to the 74.7 % reported by Scheitle et al. [239], who measured the HTTPS capabilities of the Alexa top 1 million websites. The average increase was +1.9 percentage points in a month-by-month comparison. Statistically significant changes in the variance (ANOVA) were found from December 2017 to January 2018 (+2.9), early May to June (+3.9), and October to November 2018 (+2.7). The high increase from May to June was preceded and followed by months of less increase, which can be interpreted as a concentration of activities around the GDPR enforcement date that followed an overall trend. On the top-level domain (TLD) level, the majority (18 out of 28) had adoption rates higher than 80 % by November 2018. For three countries (`.pl`, `.gr`, `.es`), we found an increase of more than 30 percentage points, but only for `.es` the adoption rate was above the average in our last measurement.

Our findings indicate that at the time the GDPR came into force websites' data processing practices, as indicated by the use of third-party tracking and HTTPS, appear largely unchanged. Next, we focus on a second development, an observed increase in the number of websites that present visitors with cookie consent notices, which, in principle, are supposed to not only inform but also

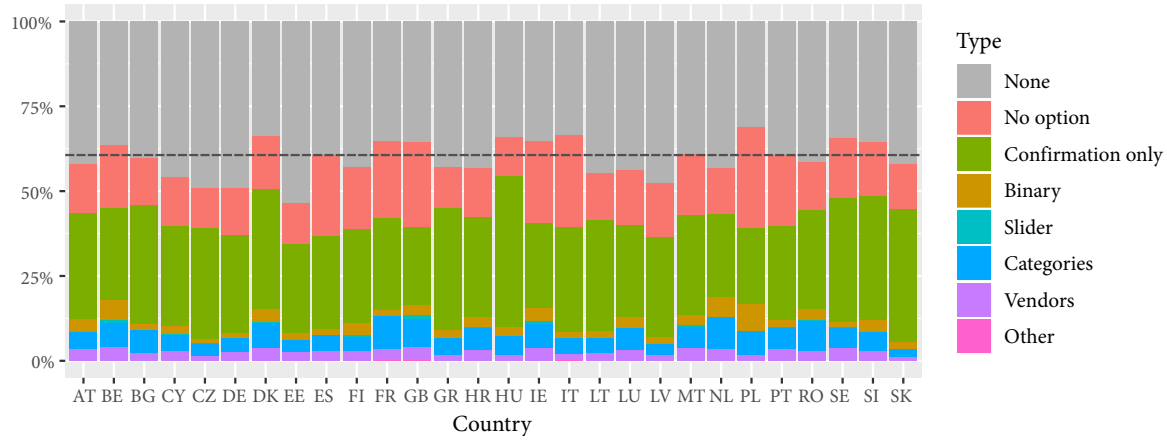


Figure 3.4: Types of consent notices by country (October 2018). The dotted line indicates the average prevalence across all countries.

provide visitors with actual choices regarding the processing of their personal data.

3.4.2 *Consent Notices*

Our manual website inspections confirmed the perceived increase in the prevalence of consent notices in the months leading up to the GDPR enforcement date. We found that the adoption of cookie consent notices had increased across Europe, from 46.1 % in January 2018 to 62.1 % at the end of May 2018 (post-GDPR). Adoption rates differ between individual member states, as does the distribution of different types of consent notices. The libraries we encountered on popular websites did not always support important features to fulfill the requirements of EU privacy law outlined in Sections 2.2.1.3 and 2.2.2, like purpose-based selection of cookies ("specific" requirement in Article 4(11) GDPR) and consent withdrawal (Article 7(3) GDPR).

3.4.2.1 *Adoption*

Consent notice adoption over time

Table 3.1 compares the prevalence of cookie consent notices in January 2018 and May 2018. Grouped by Alexa country ranking, the percentage of sites featuring a consent notice, on average, has increased, ranging from +20.2 percentage points in Slovenia to +45.4 in Italy. Looking at the sites by TLD, the average adoption rate increased from 50.3 % to 69.9 % post-GDPR. For the `.nl` and `.si` TLDs, the number of websites with consent notices did not increase substantially from January to May 2018, as these countries both already had high adoption rates of 85.2 % and 75.8 %, respectively. The highest increase in consent notice prevalence by TLD was observed in Ireland – for the 104 `.ie` domains in our data set, the adoption rate increased from 17.3 % to 87.5 %.

Notice types by country

Figure 3.4 shows the distribution of the different types of cookie consent notices (see Section 3.3.4) by country at the end of May 2018, i. e., shortly after the GDPR enforcement date. The use of category-based cookie consent notices stood out in France and Slovenia, while websites in Poland used the

Table 3.1: Prevalence of consent notices in the top 500 websites by country (Alexa website ranking) and TLD, pre- (January 2018) and post-GDPR (end of May 2018).

	Country ranking					TLD			
	n	pre %	post %	diff %		n	pre %	post %	diff %
AT	455	33.0	55.2	22.2	.at	132	45.5	69.7	24.2
BE	460	40.9	61.1	20.2	.be	141	59.6	78.7	19.1
BG	451	37.9	60.5	22.6	.bg	166	52.4	71.7	19.3
CY	432	26.4	50.2	23.8	.cy	58	13.8	27.6	13.8
CZ	459	34.0	52.7	18.7	.cz	251	44.6	58.2	13.5
DK	447	41.2	68.9	27.7	.dk	174	72.4	87.4	14.9
DE	455	26.2	49.0	22.9	.de	172	42.4	64.5	22.1
EE	441	9.5	35.8	26.3	.ee	132	14.4	35.6	21.2
ES	429	41.5	64.3	22.8	.es	86	72.1	84.9	12.8
FI	462	27.5	53.9	26.4	.fi	145	37.9	55.9	17.9
FR	453	49.2	66.9	17.7	.fr	139	77.0	87.1	10.1
GB	463	37.4	67.0	29.6	.uk	108	58.3	82.4	24.1
GR	443	40.0	59.8	19.9	.gr	233	56.7	69.1	12.4
IE	447	21.3	64.2	43.0	.ie	104	17.3	87.5	70.2
IT	423	21.3	66.7	45.4	.it	174	30.5	90.8	60.3
HU	440	46.4	62.7	16.4	.hu	228	67.1	76.3	9.2
HR	430	28.6	54.7	26.0	.hr	141	48.9	70.9	22.0
LV	434	16.8	41.9	25.1	.lv	126	38.1	61.1	23.0
LT	452	27.0	47.3	20.4	.lt	174	50.0	63.2	13.2
LU	440	24.8	51.8	27.0	.lu	61	36.1	57.4	21.3
MT	446	25.8	58.1	32.3	.mt	46	21.7	43.5	21.7
NL	459	37.3	54.2	17.0	.nl	115	85.2	87.8	2.6
PL	462	53.9	68.6	14.7	.pl	256	75.4	83.2	7.8
PT	430	31.4	53.7	22.3	.pt	116	52.6	65.5	12.9
RO	434	30.2	53.5	23.3	.ro	160	52.5	73.1	20.6
SE	459	33.3	63.6	30.3	.se	166	50.6	78.3	27.7
SK	438	42.2	56.8	14.6	.sk	189	60.3	69.3	9.0
SI	451	43.9	64.1	20.2	.si	132	75.8	77.3	1.5
Total	6,357	46.1	62.1	16.0		4,125	50.3	69.9	19.6
					.com	1,915	28.7	50.7	22.0
					.net	248	25.4	35.5	10.1
					.ru	148	5.4	6.7	1.3
					.org	119	13.5	23.5	10.8
					.eu	43	23.3	37.2	13.9
					.tr	32	6.3	6.3	0.0

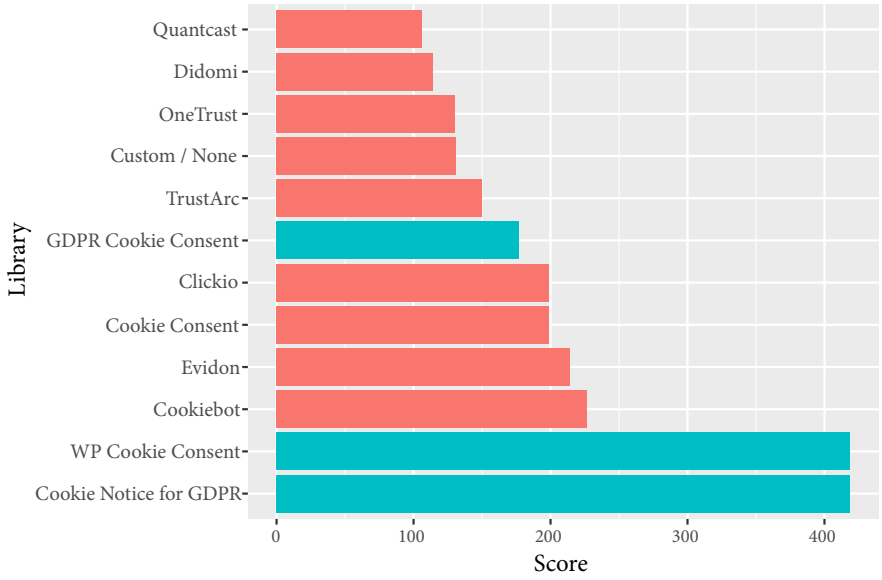


Figure 3.5: Distribution of cookie consent libraries based on the websites' Alexa rank (December 2018). The blue-green bars indicate Wordpress-only libraries.

highest number of no-option notices. Across all countries, we found fewer than 20 % of websites to provide visitors with the option to decline the use of cookies and similar tracking technology, while the vast majority did not have a consent notice, followed by sites that used a confirmation-only notice and those who only showed a no-option notice and assumed consent through visitors' continued use of the website. Neither of these variants fulfill the consent requirements of EU privacy law outlined in Sections 2.2.1.3 and 2.2.2 [236].

3.4.2.2 *Consent Libraries*

Even if the user interface of a consent notice provides options to decline the use of cookies or collection of other personal data, this does not automatically mean that the underlying website adheres to the selection made in the user interface. In fact, our analysis of consent libraries revealed that many implementations of consent notices lacked the required backend to adjust the website's data processing to the visitor's selection.

PREVALENCE We identified 31 cookie consent libraries in our data set and measured their distribution with automated means in July and December 2018. For the July measurement, we found that 15.4 % of the websites displaying cookie consent notices used one of the identified libraries. Figure 3.5 displays the scores we computed for the different libraries to get an impression of their reach (see Section 3.3.5). From our subsequent in-depth analysis we excluded two libraries not available in English and a WordPress plugin that had been discontinued in November 2018.

PROPERTIES Our results of the analysis of 28 cookie consent libraries are presented in Table 3.2. We analyzed and compared the libraries with respect to the following properties:

First- vs. third-party hosting

Source identifies whether the code for the consent notice can be hosted by the first party (*self-hosted*) or whether it is retrieved from a *third party*.

Local, decentralized, and centralized consent management

Mechanism refers to the three distinct mechanisms for consent management. One solution is to have the website asking for consent implement the (un)blocking of cookies according to the visitor's wishes (*local consent management*). The consent information is stored in a first-party cookie the website can query to react accordingly. *Decentralized consent management* leverages the opt-out APIs provided by third parties, such as online advertisers, to tell them the visitor's preferences and they are expected to react accordingly. They may remember the visitor's decision by setting a third-party opt-out cookie. A third option is to use the services of a third party offering *centralized consent management*, who is informed of the visitor's cookie preferences and triggers the corresponding notifications to participating vendors that would like to set cookies on the visitor's system. The libraries in our data set that follow this approach have implemented IAB Europe's Transparency and Consent Framework (see Section 2.2.1.3). Finally, libraries that do not provide any type of consent management are only capable of displaying a cookie notice without an underlying backend.

User interface: modality & options

Consent notices are presented to website visitors in one of two ways: *Overlays* block usage of the website until the visitor clicks one of the buttons on the notice. In contrast, *standard banners* are non-modal and thus do not prevent use of the underlying website while the notice is displayed. Regarding the options the interface may offer to website visitors, we use the same definitions as in our analysis in Section 3.4.2.1.

Technical details

The backend of any implementation of consent notices is essential to ensure that the website adheres to the visitor's selection. *AutoAccept* refers to mechanisms that automatically assume the visitor to consent to the use of cookies if they scroll or click a link on the website and react by removing the banner. This violates the requirements of EU privacy law that only an explicit, unambiguous action can constitute consent (see Section 2.2.2).

The following two properties are crucial for a library's ability to comply with the visitor's cookie preferences. The first is the ability to *block cookies*[3], i. e., prevent the website from setting cookies if the visitor has not (yet) consented to their use. If the visitor changes settings for previously set cookies, the library is expected to *delete cookies*.

Custom expiration refers to the site administrator being able to manually set the expiration date of the cookie and thus determine when the consent notice will be shown again. *Geolocation* functionality allows to display the consent notice only to visitors from selected regions, typically those where EU privacy laws apply. Finally, some consent libraries offer the website owner to automatically *scan* their site for cookies to assist with sorting them into categories or just display them to provide additional information to the visitor.

3 For the rest of this section, when we talk about *cookies* in the context of consent, we only refer to cookies that are not considered strictly necessary according to Article 5(3) of the ePrivacy Directive (see Section 2.2.2) and, thus, can only be set with the website visitor's consent.

Legal compliance

The *Legal* section lists two properties Article 7 GDPR considers vital for valid consent: the necessity for a data processor to prove that consent was given and the ability for a data subject to withdraw consent. If a library allows website visitors to reconsider and modify their previous consent by displaying a small button or ribbon that opens the consent interface again, we captured this via the *consent change* property. *Consent logging* lets the website owner store information about visitors' consent decisions for auditing purposes.

Table 3.2: Properties of cookie consent libraries as of August 2018. ●: supports this property, ○: does not support this property, B (for "bug"): functionality exists but did not work, ?: could not be determined, $: paid version only. * indicates a library we could not install on our test website. W: also available as a WordPress plugin.

		Source		Mechanism			User Interface								Technical Details						Legal	
	Reference	Self-hosted	Third party	Local CM	Decentralized	Centralized	Banner	Overlay	No Option	Confirmation	Binary	Slider	Categories	Vendors	AutoAccept	Block Cookies	Delete Cookies	Cookie Scan	Custom Expir.	Geolocation	Reevaluation	Logging
General Libraries																						
Civic Cookie Control[W]	[35]	○	●	●	●	○	●	○	○	○	●	○	●	○	○	●	●	○	●	$	●	●
Clickio Consent Tool*	[38]	○	●	○	○	●	●	●	○	●	●	○	?	●	●	○	○	?	?	●	●	●
consentmanager.net[W]	[138]	○	●	●	?	●	●	●	○	●	●	○	●	●	●	●	○	●	●	●	●	●
cookieBAR[W]	[276]	●	●	●	○	○	●	●	○	●	●	○	○	○	●	○	●	○	●	●	○	○
Cookiebot[W]	[47]	○	●	●	○	●	●	●	○	●	●	○	●	○	●	●	●	●	●	$	●	●
Cookie Consent	[248]	●	●	●	○	○	●	○	●	●	●	○	○	○	●	●	○	○	●	●	●	○
Cookie Information*	[42]	○	●	●	○	●	●	●	○	●	●	○	●	○	?	●	?	●	?	?	●	?
Cookie Script*	[203]	$	●	●	○	○	●	●	●	●	●	○	○	○	●	●	?	○	●	$	●	$
Crownpeak (Evidon)*	[79]	○	●	○	●	●	●	●	●	○	●	○	●	●	●	●	○	●	○	●	○	●
Didomi*	[54]	○	●	○	●	●	●	●	○	●	●	○	●	●	○	●	?	○	?	●	●	●
jquery.cookieBar	[296]	●	○	○	○	○	●	○	●	●	○	○	○	○	○	○	○	○	○	○	○	○
OneTrust*	[204]	○	●	●	○	●	●	?	●	●	○	○	●	●	●	●	●	●	○	●	●	●
Quantcast Choice[W]	[220]	○	●	○	○	●	●	●	○	●	●	○	●	●	○	○	○	○	●	●	●	●
TrustArc (TRUSTe)*	[277]	○	●	○	●	●	●	●	●	●	○	●	●	●	●	●	○	●	○	●	●	●
WordPress Plugins																						
Cookie Bar	[21]	●	○	○	○	○	●	○	○	●	○	○	○	○	○	○	○	○	○	○	○	○
Cookie Consent	[32]	●	○	○	○	○	●	○	●	●	○	○	○	○	●	○	○	○	●	●	○	○
Cookie Law Bar	[293]	●	○	○	○	○	●	○	●	●	○	○	○	○	○	○	○	○	○	○	○	○
Cookie Notice for GDPR	[53]	●	○	●	○	○	●	○	○	●	●	○	○	○	●	●	○	○	●	○	●	○
Custom Cookie Message	[8]	●	○	○	○	○	●	○	○	●	○	○	●	○	○	○	○	○	●	○	○	○
EU Cookie Law	[186]	●	○	●	○	○	●	○	○	●	○	○	○	○	●	●	●	○	●	○	●	○
GDPR Cookie Compliance	[185]	●	○	●	○	○	●	$	○	●	○	○	●	○	$	●	○	○	○	○	●	○
GDPR Cookie Consent	[290]	●	○	●	○	○	●	●	●	●	●	○	$	○	●	●	●	?	○	$	●	$
GDPR Tools	[221]	●	○	●	○	○	●	○	○	●	●	○	●	●	○	$	?	○	$	○	○	?
WF Cookie Consent	[298]	●	○	○	○	○	●	○	○	●	○	○	○	○	○	○	○	○	○	○	○	○
Drupal Modules																						
Cookie Control	[29]	●	○	●	○	○	●	○	●	○	●	○	○	○	●	●	○	○	○	B	●	○
EU Cookie Compliance	[208]	●	○	●	○	○	●	○	○	●	●	○	○	○	●	●	?	○	●	●	●	●
Simple Cookie Compliance	[255]	●	○	○	○	○	●	○	○	●	○	○	○	○	○	○	○	○	●	○	○	○

CONSENT MECHANISMS Combining the different types of user interfaces with the ability to block and delete cookies allows for the implementation of different consent mechanisms. Note that while we use the wording "consent" here as an umbrella term for different concepts to signal agreement (as in the "notice and consent" principle [297]), this does not automatically mean that the mechanisms described below constitute valid consent under the GDPR and ePrivacy Directive.

- *Implied Consent* mechanisms assume the visitor agrees to the use of cookies if they continue to use the website. Implementing this just requires displaying a banner with or without a confirmation button; AutoAccept may also be used. Again, note that the concept of "implied consent" does not meet the requirements outlined in Article 7 of the GDPR (see Section 2.2.1) and Article 5(3) of the ePrivacy Directive (see Section 2.2.2).
- If a site displays a notice that prevents the visitor from accessing the site unless the use of cookies is acknowledged, this is referred to as *forced opt-in*. This requires support of the overlay banner type to block access to the website and a confirmation button.
- An *opt-in* mechanism does not set any non-essential cookies by default, but visitors have the opportunity to explicitly allow the use of all the website's cookies. This requires a banner with one (allow) or two (allow / disallow) buttons that blocks cookies by default. GDPR- and ePrivacy-compliant consent needs to be "freely given" and, thus, requires an option to deny consent [236].
- In the *opt-out* case, all cookies are set by default, but the website visitor can opt out. This requires the library to display a banner with one (disallow) or two (disallow / allow) buttons and delete cookies that have already been set.
- Mechanisms with more options (slider, checkboxes, individual vendors) just require the library to implement more fine-grained deletion and blocking of cookies. Giving the website visitor more control of which types of cookies to allow and to refuse is in alignment with the GDPR's requirement that consent be *specific*, i. e., given with regard to a specific purpose [236]. It is questionable whether slider-based mechanisms are compliant with the GDPR requirement of "freely given" consent because they force the website visitor to also allow the previous categories in the hierarchy.

Examining the libraries listed in Table 3.2, we made the following observations:

Implied and forced "consent"

The notion of implied consent – widespread at the time of our measurements yet not valid consent under the GDPR – is widely supported and easy to implement: A banner that merely informs visitors about the website's use of cookies just requires adding a JavaScript library to the website or activating a WordPress plugin. The same applies to forced consent. In contrast, mechanisms that offer multiple options require more effort because whether cookies are set and read or not depends on user selection.

Opt-in

The opt-in scenario can be implemented (a) by overwriting the JavaScript object `document.cookie` and adding a conditional block that only executes when querying the consent cookie returns that the visitor has consented. We also found libraries that (b) trigger a JavaScript event when the visitor has consented, upon which the cookie-setting code is run.

Opt-out

Implementing an opt-out is challenging because it requires the cookie consent library to trigger deletion of the cookies that have already been set. A website can easily delete cookies originating from its own domain – unless they are `HttpOnly` or `Secure` cookies. It cannot delete third-party cookies due to the Same-Origin Policy (SOP) preventing access to cookies set by another host. Working opt-out mechanisms we found in the (b) scenario use JavaScript events to learn when consent has been revoked for all or selected categories of cookies and then leverage third-party opt-out mechanisms to delete these cookies. Google Analytics, for example, can be triggered to remove its cookies by setting `window['ga-disable-UA-XXXXXX-Y'] = true`, where `UA-XXXXXX-Y` references the website ID. This mechanism requires third parties to provide APIs for opt-outs. In case the third party does not, the website visitor is ideally alerted that their opt-out (partially) failed, as demonstrated by Civic Cookie Control, which displayed a warning message that the cookies could be deleted automatically and provided a link to the third party's opt-out website. This also poses limitations for cookie settings interfaces: Once a visitor has agreed to the use of third-party cookies, revoking consent is limited to cookies for which deletion can be triggered by the website.

Categorizing cookies

If a library supports consent for different cookie categories, it needs to know which cookies should be considered "strictly necessary" such that Article 5(3) of the ePrivacy Directive applies and consent is not required. If the mapping of cookies into categories is done by the website owner, nothing prevents them from declaring all cookies "strictly necessary." We found one notable example on the website of a major US TV network, where cookies for Google Analytics and Google Ad Serving were categorized as necessary for website operation. One online marketing website used a complex consent solution but had simply declared all cookies necessary, causing the library to merely display a "no option" solution.

Long lists of third-party vendors

Fine-grained consent for individual vendors is supported by libraries that implement the IAB Europe Transparency and Consent Framework. The TCF-based consent notices we encountered both provided too much and too little information: By default, the TCF's vendor-based cookie selection mechanism displays all of the vendors participating in the framework, not just the ones used by the website.[4] This renders the fine-grained control offered by the framework unusable. We drew from our dataset a sample of 24 websites with TCF-supporting consent notices (10 Didomi, 7 Clickio, 7 Quantcast) and found that only two sites using Didomi had customized their list of vendors, reducing their number to 21 and 8. At the same time, the functionality of TCF-based consent notices at the time of analysis was limited to TCF vendors, unless the library also supported other vendors as in Didomi's consent mechanism, which had integrated additional vendors including Google and Facebook. As we

4 As of December 13, 2018, the TCF supported 460 vendors [126]; as of June 11, 2022, this number had increased to 800 [128].

observed during the manual annotation of consent notices, TCF notices tended to display a standard text that did not inform visitors that the website may also use other third parties in addition to the listed TCF vendors and that those other parties are not bound by the visitor's consent decision made in the TCF-based consent plugin.

Technical challenges to GDPR-compliant consent

Our analysis shows that implementing GDPR consent requirements in practice is a challenge. Consent under GDPR requires an affirmative action by the website visitor, upon having been provided with sufficient information about the purposes of cookie use. This is at odds with usability, as prior work has shown the ineffectiveness of previous choice mechanisms [154]. The options to implement meaningful choices for website visitors, including the ability to withdraw consent, are limited by technical restrictions, such as the SOP, a core principle of web security, and the business interests of third parties, not all of which are interested in providing an opt-out API, let alone GDPR-mandated opt-in mechanisms. Under the GDPR, consent has to be given for specific purposes of data processing, which raises the question who defines the purpose of the use of a certain cookie. If left to the developers or website owners, it is prone to abuse of the "strictly necessary" category to circumvent the consent requirement in the ePrivacy Directive.

3.5 DISCUSSION

Our results show that at the time the GDPR went into effect websites made changes, mostly transparency-related, that can be considered improvements for visitor privacy, but most still do not meet GDPR standards regarding the implementation of valid consent and the "data protection by default" principle. We discuss resulting challenges and opportunities for researchers, policymakers, and companies. We also discuss some limitations of our study.

3.5.1 *Impact of the GDPR*

Our analysis focuses on the 28 EU member states as of 2018, but the GDPR also impacts websites from other countries – first, because some non-EU countries have decided to adopt similar rules (e. g., Norway, Switzerland, Iceland and Liechtenstein [74]), and second, because websites that offer services in the EU have to comply with the GDPR. For example, according to the Alexa ranking, 53 % of the top 500 websites in the US and 48 % of the most popular sites in Russia also appear in at least one EU state's list of the 500 most popular websites.

An encouraging finding of our study is that even though many implementations of cookie consent notices are insufficient to obtain informed consent under the GDPR and ePrivacy Directive, the prevalence of such notices increased by 16 % around the GDPR enforcement date. This suggests that websites felt compelled to at least make *some* changes with regard to how they inform visitors about the site's data processing practices. The fact that increased prevalence of consent notices could be observed across all EU member states suggests that in principle the harmonization of data protection rules is a suitable tool to increase privacy protection across Europe. However, despite this trend, actions

taken to comply with the GDPR vary greatly, partially due to national DPAs issuing diverging advice on the concrete implementation of requirements from EU privacy law [236].

3.5.2 *Need for More Detailed and Practical GDPR Guidance*

Although the GDPR makes it clear that websites need to be transparent about their data processing practices, details about what is permissible or required remain unclear. With respect to cookie consent notices, the observed variance in implementation indicates the need for clearer guidelines for website owners. Such guidance should, for example, clarify which types of cookies can be set on what legal grounds. This requires decisions on questions such as whether website operators can claim a "legitimate interest" in web analytics or if user tracking requires explicit consent. There is still hope that a future ePrivacy Regulation may provide some clarity regarding these issues, but when we conducted this study in 2018, it was unclear when and it what form it may be adopted, and this still holds four years later at the time of writing this thesis. Our results also show that some countries lagged behind in the adoption of cookie consent notices. To improve the situation, data protection authorities could support website operators by providing effective guidelines and other tools for cookie handling, consent mechanisms, and privacy statements.

3.5.3 *False Sense of Compliance*

Some of the uncertainty about how to interpret the GDPR may have resulted in a false sense of compliance. Although the prevalence of consent notices increased by 16.0 % between our pre- and post-GDPR measurements, 37.9 % of websites still did not have one in late May 2018, and the vast majority of consent notices did not provide website visitors with options to decline the use of cookies and similar tracking mechanisms. Since our study was conducted, consent notices that implement IAB Europe's TCF have gained wider traction [114]. These notices typically allow for at least a binary selection but, as we will see shortly in Chapter 4, frequently employ *dark patterns*, deceptive interface designs that try to steer people towards a specific action, in this case consenting to all data collection [195]. This is at odds with the GDPR requirement that it is as easy to deny consent as it is to give it [236]. Even if notices' user interfaces provide sufficient choice, the underlying backend often does not properly implement user selection, causing cookies to be set and visitor data to be collected before or without valid consent [174], as we will investigate more closely in Chapter 6. Thus, website owners need to me made aware of the fact that many widely used implementations and/or configurations of cookie consent are not sufficient to obtain visitors' valid consent under the GDPR and ePrivacy Directive. In the case of "implied consent," the A29DPWP explicitly stated that *"merely proceeding with a service cannot be regarded as an active indication of choice*" [14], and in 2019 the ECJ declared the use of pre-ticked checkboxes to not constitute valid consent [69]. Similar binding clarification could be in order for other frequently disregarded requirements of consent, including hiding rejection options under a second layer or highlighting the "Accept all" option with color.

Websites could be motivated to implement such guidelines by making them aware of the GDPR's instrument of fines (see Section 2.2.1.6), which is a question we will investigate in Chapter 6.

3.5.4 *Opportunities for Web Privacy and Security Research*

The mere presence of a cookie consent notice does not mean that a website is compliant with privacy law. In our work we used manual inspection to find and classify consent notices on websites, which required significant time and human resources and, thus, is not feasible at scale and over time. The differences in popular implementations we identified make it difficult to automatically detect and analyze consent notices at scale. Consequently, subsequent web privacy measurement studies have focused on notices that implement the IAB Europe TCF or even specific CMPs within that framework [19, 114, 174, 195, 197]. In Chapter 6 we introduce a vendor-independent approach to detect consent notices, but it is designed to keep the number of false positives low without considering false negative rates. Future work could refine existing approaches and develop more accurate detection mechanisms that aid future web privacy measurements of consent notices and help inform regulators in their enforcement of privacy legislation. Going further, web privacy research could develop mechanisms to automatically assess whether a consent notice is compliant with EU privacy law. As with automated evaluation of privacy policies, this is a complex challenge that could require deeper analysis of web applications to identify click paths and Natural Language Processing (NLP) in multiple languages to evaluate notice text.

3.6 CONCLUSION

Our analysis of the 500 most popular websites in each of the EU member states identified first positive effects on web privacy that occurred around the GDPR enforcement date. Most notable is the increased prevalence of cookie consent notices, which now greet European Web users on more than half of all websites. While seemingly positive, the increase in transparency may lead to a false sense of privacy and security for website visitors. Few websites offered their visitors actual choice regarding the use of cookies and tracking through third parties. Moreover, most of the analyzed cookie consent libraries did not meet GDPR requirements, an observation confirmed by subsequent measurement studies.

Browser manufacturers and the online advertising industry so far have not been able to agree on technical privacy standards, while previous standards such as DNT failed due to their non-binding nature. This puts an additional burden on website visitors, who are presented with an increasing number of privacy notifications that, even if they fulfill the law's transparency requirements, are unlikely to actually help Web users make more informed decisions regarding their privacy. In the absence of self-regulation in the industry, public regulators need to provide clear guidelines in what cookies a service can claim "legitimate interest" and which require actual consent.

4

WEBSITE VISITORS' INTERACTION WITH CONSENT NOTICES

4.1 INTRODUCTION

Lack of real change, but increased transparency

On a high level, our assessment of changes on European websites around the GDPR enforcement date presented in Chapter 3 led to two observations: While the number of third-party services and cookies on European websites barely changed – an observation shared by concurrent related work [251] –, websites have evolved to prompt visitors for consent to the use of cookies and similar tracking technology. At the end of May 2018, we found about 62 % of popular websites in the EU to display a consent notice, an increase of 16 % compared to our measurement in January, with some countries having an observed increase of up to 45 percentage points.

Increase in complexity but no real control

During manual annotation of the data set we noticed that the design and complexity of such consent notices greatly varied: The majority of notices merely stated that the website used cookies, providing only a confirmation button or assuming consent through continued use of the website. On the other end of the complexity spectrum, we found notices implementing the IAB Europe TCF that asked website visitors for consent to data collection for different purposes by up too 400 listed third-party vendors. Beyond overwhelming people with this high number of options, related work conducted after the study presented in this chapter found that the notices supporting this framework are often not properly implemented [174] and use dark patterns to nudge visitors into giving consent [195].

Consent fatigue and workarounds

Paired with the fact that consent notices often cover parts of the website's main content or even block access until a decision is made, this development has led website visitors to become fatigued with consent mechanisms [25]. Consequently, tools have emerged that provide pragmatic workarounds – one example is the "I don't care about cookies" browser extension [143] that tries to automatically hide consent notices. But oftentimes this only leads to data collection taking place without consent, since the default on many websites is to employ user tracking *unless* the visitor has opted out [89], and in our work in the previous chapter we found about 80 % of popular EU websites not to offer any option for refusal at all.

Design space only partially explored

Our verdict from these developments was that consent notices had become ubiquitous but most provided too few or too many options, leaving people with the impression that their choices are not meaningful and fueling the habit to click any interaction element that causes the notice to go away – instead of actively engaging with it and making an informed choice, as intended by the GDPR and ePrivacy Directive. Still, we have also seen some notices that made better use of the available design space and, for example, did not force visitors to accept cookies, asked for consent without hidden pre-selections, or provided visitors with granular yet easy-to-grasp mechanisms to control the website's data processing practices. Thus, we hypothesized that how a consent notice

asks for consent has significant impact on website visitors' interaction with it, and that there are design decisions that better motivate people to interact with consent notices in a meaningful way instead of annoying them.

Our contribution: Design space and field study of consent notices

In this chapter, we investigate this hypothesis and conduct the first field study of consent notices on a live website to systematically evaluate design properties of consent notices and their effects on consent behavior. For this, we first systematize consent notices using a sample of 1,000 notices collected from live websites and identify common variables of their user interfaces. Our research goal is to explore the design space for the user interface of consent notices to learn how to encourage website visitors to interact with a notice and make an active, meaningful choice. Over the course of four months, we conduct a between-subjects study with 82,890 real website visitors of a German e-commerce website and investigate their (non-)interaction with different variants of consent notices. We collect passive clickstream data to determine how website visitors interact with consent notices and invite them to participate in a voluntary follow-up online survey to obtain qualitative feedback. The study comprises three distinct field experiments to answer the following research questions:

- Does the position of a cookie consent notice on a website influence website visitors' consent decisions? (Experiment 1, n = 14,135)
- Do the number of choices and nudging via button formatting and preselections influence website visitors' decisions when facing cookie consent notices? (Experiment 2, n = 36,530)
- Do the presence of a privacy policy link or the use of technical / non-technical language ("this website uses cookies" vs. "this website collects your data") influence website visitors' consent decisions? (Experiment 3, n = 32,225)

In a short follow-up survey answered by more than 100 participants, we ask website visitors to voluntarily report the motivation for their selection, how they perceive the notice they have seen, and how they expect consent notices to function in general.

Influence of websites' layout

In Experiment 1 we find that visitors are most likely to interact with consent notices placed at the bottom (left) position in the browser window, while bars at the top of the screen yielded the lowest interaction rates. This is mainly due to the (un)importance of the website content obstructed by the notices and suggests taking into account characteristics of the individual website to identify the notice position most likely to encourage user interaction.

Limited number of options

Experiment 2 finds higher interaction rates with notices that provided at most two options, compared to those that let visitors (de)activate data collection for different purposes or third-party services individually, even though those notices do not allow visitors to express consent freely. We also show that the more choices are offered in a notice, the more likely visitors were to decline the use of cookies. This underlines the importance of finding the right balance between providing enough detail to make people aware of a website's data collection practices and not overwhelming them with too many options. At

the same time, nudging visitors to accept privacy-invasive defaults leads more visitors to accept cookies, whereas in a privacy-by-default (opt-in) setting, less than 0.1 % of visitors allowed cookies to be set for all purposes. This suggests that the current data-driven business models of many webservices, who often employ dark patterns to make people consent to data collection, may no longer be sustainable if the GDPR's "data protection by design and by default" principle is enforced.

Limited influence of language and privacy policy link

Experiment 3 shows that technical language ("This site uses cookies" instead of "This site collects your data") appears to yield higher interaction rates with the consent notice but decreases the chance that website visitors allow cookie use. We find that the presence of a link to the site's privacy policy does not increase visitor interaction, underlining the importance of making information immediately actionable rather than pointing to further resources.

Visitors' preferences

Survey feedback indicates that website visitors favor category-based choices over a vendor-based approach, and they expressed a desire for a transparent mechanism. A common motivation to give consent is the assumption that the website cannot be accessed otherwise.

Recommendations

Based on the results of our field study, we conclude that opt-out consent notices are unlikely to produce intentional or meaningful consent expression. Therefore, we recommend that websites offer opt-in notices based on categories of purposes. Above all, we observed that the majority of website visitors do not accept cookies for all purposes, and feedback from our survey suggests that a unified solution that does not interfere with every single website yet provides more control than a simple yes–no decision would best fit users' needs.

4.2 RELATED WORK

In Sections 2.2.1.3 and 2.2.2 we already established the GDPR and the ePrivacy Directive as the legal frameworks that prompted websites to adopt consent notices. In this section we provide an overview of related work in the area, before we proceed to identify the design space for consent notices in Section 4.3.

4.2.1 *Consent Notices*

Multiple measurement studies of varying scope have provided insights into the prevalence of consent notices [13, 284], including our work presented in Chapter 3. Even though we have seen that many consent notice libraries can be configured to only display a notice to EU visitors, a website's top-level domain (TLD) was found to be the primary factor in whether a consent notice is displayed, rather than a visitor's location [284].

Backend

Sanchez-Rola et al. [234] evaluated the functionality of consent notices and opt-out mechanisms under the GDPR. They manually visited 2,000 popular websites, tried to opt out of data collection whenever possible, and studied the effects on the website's cookies. They found that 92 % of websites set at least one high-entropy cookie before showing any kind of notice. Only 4 % of notices provided an opt-out choice, and 2.5 % of websites removed some cookies upon opt-out. Our work presented in Chapter 3 further found that many third-party consent libraries either lack the functionality to block or

delete cookies, or require significant modification of a website to properly react to visitors' consent choices.

User interface

Prior to the work presented in this chapter, the user interface of consent notices had been classified by the provided information [147], offered choices (Sanchez-Rola et al. [234] and our work in Chapter 3), and if the notice blocks access to the website [234]. Van Eijk et al. [284] reported some statistics on the height and width of consent notices, their location offset, and notices' word and link/button counts. In Section 4.3, we systematically approach this topic and present a detailed analysis of variants in consent notices' user interfaces.

Kulyk et al. [147] investigated website visitors' perceptions of and reactions to differently worded cookie consent notices. They identified five categories of disclaimers based on the amount of information provided about the purposes of cookie use and the parties involved. In a qualitative user study, they found that the text of a cookie notice does not significantly influence Web users' decisions to continue using a website; their decision was rather based on the website's perceived trustworthiness and relevance. The participants perceived cookie consent notices as a nuisance or threat to their privacy, and they reported to lack information about the implications of cookies and possible countermeasures.

Schaub et al. explored the design space for privacy notices and controls, including consent notices and permission prompts on mobile devices [238].

4.2.2 *Perception of Cookie Control Mechanisms*

Prior work has also studied how website visitors perceive cookies and mechanisms to control their use by websites. Using Dutch panel data, Boerman et al. [18] explored how users protect their online privacy. Given the opportunity to decline cookies, many participants self-reported that they decline cookies "often" (16 %) or "very often" (17 %). Facing the decision to either accept cookies or leave the website, 12 % and 13 % reported to refrain from using the site "often" and "very often," respectively. Ha et al. [106] studied the usability of two cookie management tools in focus groups and identified misconceptions about cookies and associated risks. Kulyk et al. [148] developed and tested a privacy-friendly cookie settings interface for the Chrome browser and found that users appreciate tools that help them better understand the standard browser cookie settings, such as an assistant that transforms users' privacy preferences into cookie settings or additional explanations about the purpose and security and privacy implications of different types of cookies. Previous work has also evaluated the usability of different tools to opt out of targeted advertising [90, 107, 154] and found that users find it difficult to locate, configure, and understand these mechanisms.

4.2.3 *UX Design for Web Notices and Warnings*

From a more general perspective, warning research and ad placement studies provide insights into the effects of user interface design choices on user attention and behavior; examples include color [247] and position [28]. Studies investigating different notice designs were conducted, for example, for SSL [81], browser security [223], and phishing warnings [62].

Mathur et al. [173] classified common dark patterns in web services. In their classification scheme the observed actions are described as "sneaking" (attempting to misrepresent user actions, or delay information that, if made available to users, they would likely object to), "misdirection" (using visuals, language, or emotion to steer users toward or away from making a particular choice), and "forced action" (forcing the user to do something additional in order to complete their task). As we will see shortly, these also are broadly used in cookie consent notices.

4.3 THE DESIGN SPACE FOR CONSENT NOTICES

In Chapter 3 we have seen that consent notices currently found on websites vary both in terms of their user interface and their underlying functionality. Regarding the latter, some are only capable of displaying a notification that the website uses cookies or collects visitor data, without providing any functionality to make the website comply with the visitor's choice. Other consent notices offer complex opt-in choices and block cookies until the user consents explicitly.

In this chapter we focus on the *user interface* of consent notices, a topic which, as we have just seen in Section 4.2, has not been systematically studied before. To this end, we now determine the design space for the user interface of consent notices. We identify variables of the user interfaces of consent notices and their possible values as commonly used in different types of consent notices found on live websites in summer 2018, shortly after the GDPR became effective.

Sample of consent notices

In order to identify common properties of notices' user interfaces, we analyzed a sample from the set of notices we had collected in August 2018 and manually classified for the study presented in Chapter 3. Back then, we had already identified six distinct types of choices consent notices offer to website visitors, as described as a reminder below. In the following, we extend our prior analysis to other variables of the user interface of consent notices. For this, we took the 5,087 consent notices collected in the second round of full manual annotation in August 2018, drew a random sample of 1,000 notices, and manually inspected how they differed in their user interface. We identified the following eight variables, whose possible values, along with their frequency in our random sample, are listed in Table 4.1:

Variables of consent notices' user interfaces

SIZE. The size of the consent notice as displayed in the browser. We found the value of this variable to vary widely depending on the implementation of the notice, from small boxes that only cover a fraction of the viewport to notices taking up the whole screen. Responsive web design may result in the same notice using up different shares of the viewport, depending on the screen size and orientation of the device used to view the website. Typically notices take up a larger percentage of the viewport on smartphones than on desktop computers and tablets. The size of a consent notice may also be fixed by design, i. e., to cover the whole viewport of any device.

POSITION. We observed the consent notices in our dataset to be displayed in seven distinct positions: in one of the four corners of the viewport

(*dialog* style; 6.9 %), at the top (27.0 %) or bottom (57.9 %) like a website header or footer (*bar* style), and vertically and horizontally centered in the middle of the viewport (7.8 %). On smartphones in portrait mode, the limited space reduces the number of options to the top, bottom, and middle of the screen.

BLOCKING. Some consent notices (7.0 %) prevent visitors from interacting with the underlying website before a decision is made [238]. The site's content may also be blurred out or dimmed [89]. All consent notices shown in the center position were blocking. We also observed some blocking consent notices at the top or bottom position.

CHOICES. Consent notices offer website visitors different choice options. In Chapter 3 we already identified the following mechanisms for user interaction:

- *No option* notices simply inform the visitor that the website uses cookies without any option for interaction. The visitor continuing to use the website is interpreted as agreement to the notice.
- *Confirmation-only* banners feature a button with an affirmative text such as "OK" or "I agree", clicking on which is interpreted as an expression of consent. Like *No option* notices, this does not provide website visitors with sufficient choice to constitute valid consent under the GDPR and ePrivacy directive [236].
- *Binary* notices provide two buttons to either accept or decline the use of all cookies on the website.
- *Category*-based notices group the website's cookies into a varying number of categories. Visitors can allow or disallow cookies for each category individually, typically by (un)checking a checkbox or toggling a switch. For transparency reasons, the category of "strictly necessary" cookies (whose use does not require consent according to Article 5(3) of the ePrivacy Directive as described in Section 2.2.2) is often also listed but the switch to deactivate it is rendered inactive. Some notices use a *slider*: Instead of (de)selecting categories individually, the user can move a slider to select one of the predefined levels, which implies consent to all of the previously listed categories.
- *Vendor*-based notices offer even more fine-grained control by allowing visitors to accept or decline cookies for each third-party service used by the website. When we devised this classification in Chapter 3, we found this type of notice in banners implementing IAB Europe's Transparency and Consent Framework [132], which refers to its advertising partners as "vendors."

TEXT. The text displayed by consent notices also varies widely. It should inform the website visitor of the fact that the website uses cookies or similar tracking technology and may list additional information such as the purpose of the data collection. Depending on the choices offered, the notice may provide instructions for consenting to (or denying) the use

of cookies. In Table 4.1 we provide an overview of common text contents of consent notices for the following typical pieces of information:

- *Collection.* What the visitor consents to, which can be the use of cookies, the collection of website visitors' personal data, both, neither, or something else (such as the website's privacy policy).
- *Processor.* Who collects this information, which can be specifically limited to the first party or third-party services, both, or refer to an unspecified party (usually denoted by the pronoun "we" or the domain or website name).
- *Purposes.* The stated purposes of data collection may be specific (e. g., "audience measurement" or "ad delivery"), generic (e. g., "to improve user experience"), or not specified at all in the notice itself.

NUDGING & DARK PATTERNS. More than half (57.4 %) of the consent notices in our sample use interface design to steer website visitors towards accepting privacy-unfriendly options. Typical techniques include color highlighting of the button to accept privacy-unfriendly defaults, hiding advanced settings behind hard-to-see links, and pre-selecting checkboxes that activate data collection [43]. We observed all of these techniques in our sample. Subsequent work conducted by Nouwens et al. [195] in September 2019 confirmed this. After the study presented in this chapter had been conducted, in October 2019, the European Court of Justice (ECJ) declared the use of preselected checkboxes to violate the ePrivacy Directive's requirements for valid consent, as the preselection does not constitute a clear, affirmative action by the visitor [69]. In the aftermath of this ruling, our own continuous observations as Web users hinted at preselected checkboxes becomming less common.

FORMATTING. We found that, unless predetermined by the consent library used, the choice of fonts and colors typically matched that of the underlying website. The formatting of consent notices may also be influenced by the website's business requirements [89], for example, sites relying on monetization via OBA are unlikely to steer their visitors towards an opt-out mechanism by making this option highly visible.

LINK TO ADDITIONAL INFORMATION. Consent notices may include a link to the website's privacy policy, a designated cookie policy, or a website providing additional information about cookies. 92.3 % of the notices in our sample contained such a link to additional information. In Table 4.1, we marked as "other" the rare case of consent notices where the full privacy policy was already included in the notice itself.

Table 4.1 shows that the majority of consent notices in our sample were placed at the bottom of the screen (58 %), not blocking the interaction with the website (93 %). They offered no options besides a confirmation button that did not do anything (86 %), and most tried to nudge users towards consenting (57 %). While nearly all notices (92 %) contained a link to a privacy policy, only a third (39 %) mentioned the specific purpose of the data collection or who could access the data (21 %).

Table 4.1: Variables of the user interface of consent notices and their values across a sample of 1,000 drawn from 5,087 consent notices collected from the most popular websites in the EU in August 2018. We did not analyze the value space for size and formatting due to the high number of possibilities. Nudging is not available (N/A) for "no option" notices.

Position		*Choices (visible)*		*Choices (hidden)*	
top	27.0 %	no option	27.8 %	no option	26.3 %
bottom	57.9 %	confirmation	68.0 %	confirmation	59.9 %
top right	0.2 %	binary	3.2 %	binary	4.0 %
bottom right	3.0 %	categories	1.0 %	slider	0.2 %
top left	0 %	vendors	0 %	categories	8.1 %
bottom left	3.7 %			vendors	1.1 %
center	7.8 %			other	0.4 %
other	0.4 %				
Blocking		*Nudging*		*Link to privacy policy*	
yes	7.0 %	yes	57.4 %	yes	92.3 %
no	93.0 %	no	14.8 %	no	6.6 %
		N/A	27.8 %	other	1.1 %
Text: Collection		*Text: Processor*		*Text: Purposes*	
"cookies"	94.8 %	unspecified	75.5 %	generic	45.5 %
"data"	1.4 %	first party	0.7 %	specific	38.6 %
both	1.6 %	third party	2.6 %	none	16.9 %
neither	0.9 %	both	21.1 %		
other	1.3 %	other	0.1 %		

4.4 METHOD

Given the requirements for explicit, freely given, informed consent, as outlined in Sections 2.2.1.3 and 2.2.2, the vast majority of cookie consent notices we analyzed are likely not compliant with European privacy law. To further investigate the effects of different combinations of the identified UI properties on consent behavior, including those that actually implement real informed consent under European privacy law, we conducted a field study with consent notices on a German e-commerce website.

Investigated UI parameters

We investigated the effect of the following parameters on visitors' interactions with consent notices:

- The *position* of the notice, as notices displayed in some parts of the screen are more likely to be ignored.
- The number of *choices* offered by the notice, which is influenced by legal requirements and the need to give visitors actual control over the website's data processing without overwhelming them with too many options.

- *Nudging* visitors towards giving consent through highlighting and preselection, since this may cause people to consent who would not have made the same decision otherwise.
- The presence of a *privacy policy link* and use of *(non-)technical language*, which we dubbed *Text: Collection* in Table 4.1, i. e., whether the notice refers to "cookie use" (technical language) or "data collection" (non-technical language). These differences in wording may influence people's expectations of the website's data processing practices and, thus, their consent decision.

Uninvestigated UI parameters

We did not evaluate the effects of the following parameters: blocking (because the owner of our partner website had asked us not to block access to the site), formatting (because of the multitude of options – we chose the same color scheme as in the notice previously used on the website), and size (which is difficult to adjust consistently across devices).

Study overview

From the end of November 2018 to mid-March 2019, we conducted three between-subjects experiments to determine if, and how, different parameters of consent notices influence interaction rates. In each experiment, we tested variants for one or two of the parameters described in Table 4.1: position in Experiment 1, choices and nudging in Experiment 2, and wording and the presence of a privacy policy link in Experiment 3. The respective other parameters were kept constant in an experiment.

4.4.1 *Study Setup*

Partner website

We partnered with a German-language e-commerce website based on WordPress. At the time of the study, the website had 15,000–20,000 unique visitors per month, most of which were single-page visitors that reached the site from a search engine looking for product information and reviews. The third-party services used by the website were Google Fonts and the CSS framework Ionic for design, Google Analytics embedded via Google Tag Manager for audience measurement, Facebook social media buttons, embedded YouTube videos, and targeted advertisements delivered by Google Ads. All of these services store cookies in the visitor's browser.

Modified consent plugin

We modified a WordPress plugin, Ginger – EU Cookie Law[1] [170], to test different notice variants. Ginger was selected because it could block cookies before opt-in, log users' consent, and because it had been released under a GPLv2 license. We added support for checkbox-based and "no option" notices. We did not implement "slider" notices because they are essentially a category-based banner with an added restriction of which combinations of categories may be selected.

Study workflow

The plugin was further modified to function as follows in our study: When a user first visited our partner website, they were shown one consent notice. Which notice of the n test conditions in the current experiment was

1 By the time of writing this thesis, the original version of the plugin had been discontinued, but its code and description can still be accessed via the Internet Archive: `https://web.archive.org/web/20190118012018/https://wordpress.org/plugins/ginger/`.

displayed was determined in round-robin fashion. The ID of the displayed notice was stored in a cookie in the participant's browser to ensure visitors who did not click the notice would continue to see the same notice across subpages and recurring visits. Each participant was assigned a unique identifier: pid=SHA-256($ip\_address||user\_agent$). The participant's IP address was discarded after computation of pid. The participant ID was stored in another cookie, together with the participant's consent as required by Article 7 GDPR[2].

If the visitor clicked any interaction element that would usually cause a consent notice to disappear, i. e., the 'X' discard button, "Accept," "Decline," or "Submit,"[3] the notice did not disappear instantly. Instead, the notice content was replaced with an invitation to take an online survey about the visitor's experiences with the displayed notice and other previously encountered consent notices (see Appendix A.2). The invitation disclosed that this was a university study and that survey participants could win one of fifteen 25-euro shopping vouchers. Website visitors could either click "Discard" to dismiss the notice or select "Participate" to open the survey in a new browser tab. The survey was implemented in a LimeSurvey instance running on a web server hosted by our institution.

If the website visitor did not interact with the consent notice, the content of the notice was automatically replaced with the survey invitation 30 seconds after the page had fully loaded. This is because we also wanted to explore website visitors' reasons for *not* interacting with consent notices. According to web analytics data for our partner website, 95 % of all visitors who had interacted with the website's original consent notice had done so within 30 seconds of accessing the site. Thus, we assumed that website visitors who did not interact with the consent notice within 30 seconds would not have clicked it at a later point in time.

Logging user interactions

We modified the Ginger plugin's logger add-on to create log entries whenever a website visitor clicked an interaction element on the notice. Log events were also triggered upon page load, when links to the privacy policy or survey were clicked, when the consent notice content was auto-replaced with the survey invitation, and when the visitor dismissed this invitation. Each log entry consisted of a timestamp, the participant's ID (pid), the ID of the consent notice they had seen, the event they had triggered, their screen resolution, operating system, browser, and whether an ad blocker had been detected[4].

4.4.2 *Experiment 1: Position*

Experiment 1 ran for 19 days, from November 30 to December 18, 2018. We had observed consent notices being shown at various screen positions and

2 The legal bases for storing the cookie that remembers the banner ID are Article 6(1)(e) GDPR (public interest in conducting this study) and Article 6(1)(c) GDPR (compliance with a legal obligation) for storing the consent cookie.

3 In all experiments, all texts in the consent notice and survey were in German to match the website's language. Survey responses were also in German. We translated all texts and responses into English for publication. Both the original and the translated consent notices and the survey questionnaire are available in our GitHub repository at `https://github.com/RUB-SysSec/uninformed-consent`.

4 We used BlockAdBlock 3.2.1 [7] to detect ad blocking functionality in the visitor's browser.

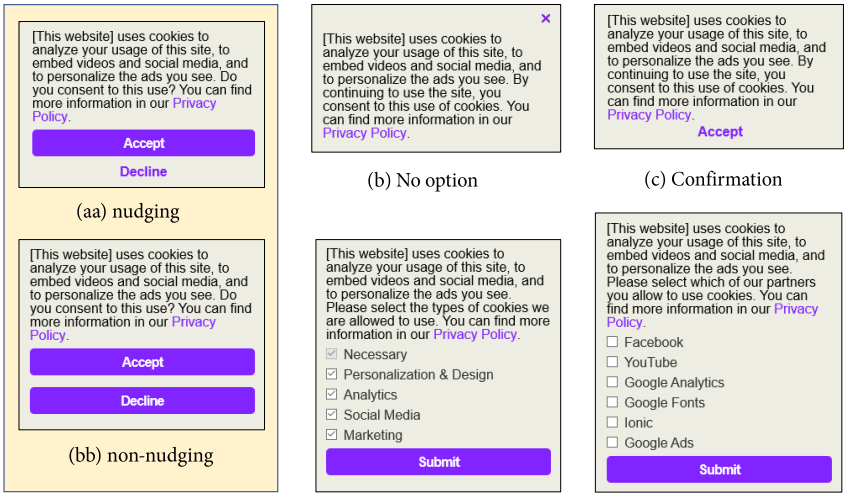


Figure 4.1: Cookie consent notices with different choice mechanisms and nudging used in our experiments: (a) a binary notice in two variants, one that color-nudges visitors to click "Accept" (aa) and one that presents both choices in the same format (bb); (b) a no-option notice (nudging not applicable); (c) a confirmation-only notice (shown without nudging); (d) a category-based notice with pre-selected checkboxes (nudging); and (e) a vendor-based notice with unchecked checkboxes (non-nudging).

wanted to determine the effect of placement on interaction with the cookie consent notice to inform our subsequent experiments. The research question for Experiment 1 was: *Does the cookie consent notice's position on a website influence a visitor's consent decision?* In order to encourage user interaction, we displayed a "binary" notice without nudging (see Figure 4.1 (a) (bb)), the simplest type that offers an actual choice. We tested the notice in six different positions (see Figure 4.2). We could not test the center position as our partner had asked us not to block access to their website.

4.4.3 *Experiment 2: Number of Choices and Nudging*

From December 19, 2018 to January 28, 2019, we conducted Experiment 2, which focused on the effects of given choices and pre-selections on consent. In our analysis of consent notices, we had identified various complexity levels of choices offered and methods to emphasize certain options. Prior work has shown that the design and architecture of choices heavily influence people's decisions [272, 291]. While this effect has also been shown successful in improving user privacy [3, 5], in practice it is most often used to make users share more information [43].

Website owners often have an interest in getting visitors to agree to the use of cookies and hence highlight certain choices in the consent notice to nudge visitors towards accepting. We observed this for 57.4 % of the notices in our sample. Therefore, our research question for Experiment 2 was: *Do the*

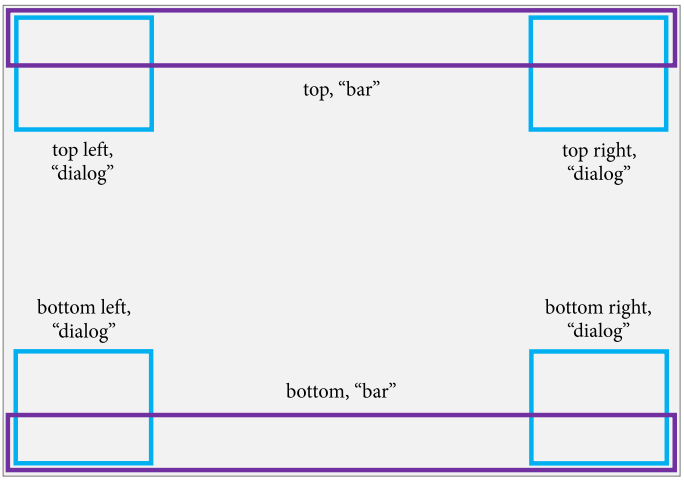


Figure 4.2: Positions tested in Experiment 1.

number of choices and nudging through emphasis or pre-selection in consent notices influence website visitors' consent decisions?

As a basis for the notices in this experiment we used the different choice mechanisms identified in Chapter 3 and listed in Section 4.3 (with exception of the slider, as mentioned earlier). For nudging, we used pre-ticked checkboxes and buttons highlighted in contrasting colors, techniques often used to nudge users towards accepting default settings [43]. While we observed that most category- and vendor-type notices in practice display such fine-grained controls only after the visitor clicked "Settings," we chose to immediately show all available options to ensure that our conditions only varied in the number and framing of choices. In Experiment 2, we displayed the following consent notices at the position determined in Experiment 1 to yield the highest interaction rates:

- *No option* (Figure 4.1 (b)): In line with many notices we observed, we added an 'X' in the top-right corner to allow website visitors to dismiss the banner. There is no nudging variant because the notice does not offer any choice.
- *Confirmation–Non-nudging* (Figure 4.1 (c)): This notice has an "Accept" button which is not highlighted.
- *Confirmation–Nudging*: Same as the Confirmation–Non-nudging notice, but the "Accept" button is highlighted (like the "Accept" button in Figure 4.1 (a) (aa)).
- *Binary–Non-nudging* (Figure 4.1 (a) (bb)): The "Accept" and "Decline" buttons have identical formatting, neither is emphasized.
- *Binary–Nudging* (Figure 4.1 (a) (aa)): Same as Binary–Non-nudging, but only the "Accept" button is highlighted in a contrasting color.

- *Categories–Non-nudging*: Same as notice (d) in Figure 4.1, but with unticked checkboxes. The "Necessary" category cannot be unchecked, as is common practice.
- *Categories–Nudging* (Figure 4.1 (d)): Same as Categories–Non-nudging, but with pre-checked checkboxes for all categories.
- *Vendors–Non-nudging* (Figure 4.1 (e)): Similar to the categories variant, but the checkboxes correspond to the third-party services used by our partner website.
- *Vendors–Nudging*: Same as Vendors–Non-nudging, but with pre-selected checkboxes.

Mapping of third parties to categories

For the category-based notices, we had to map the third-party services used by the website to different categories. We manually inspected the 434 category-based notices in our initial set of 5,087 consent notices for common category wordings. For example, we found advertising cookies to be categorized as "marketing" or "advertising"; web analytics was also referred to as "performance cookies," "statistics," or "audience measurement." This yielded the following category–third party mappings:

- *Necessary:* Cookies to remember the displayed notice and the website visitor's consent decision.
- *Personalization & Design:* Ionic, Google Fonts
- *Analytics:* Google Analytics
- *Social Media:* Facebook, YouTube
- *Marketing:* Google Ads

For all category- and vendor-based notices in Experiments 2 and 3, the available options were displayed in random order, except for the "Necessary" category, which was always displayed first, as in the majority of category-based notices we had observed on websites.

In Experiments 2 and 3, we increased the font size of the banner message, resulting in larger notices. We did this to fix an implementation bug of the Ginger plugin that had caused the text to be displayed in a very small font on some smartphones in portrait mode.

4.4.4 *Experiment 3: (Non-)Technical Language and Privacy Policy Link*

Experiment 3 was conducted from January 29 to March 15, 2019. In this experiment, we tested the influence of the presence of a link to the website's privacy policy. Previous research suggests that (American) Internet users have consistent misconceptions about privacy policies, indicated by the fact that a majority believes the existence of a privacy policy means that a website cannot share personal data with third parties [278]. At the same time, Martin [172] showed that the existence of a reference to a privacy policy in the context of data sharing explanations increases mistrust in a website. There are further

known misconceptions about what cookies actually are and what they are used for [106, 180]. To learn more about the influence of these factors in the context of consent notices, our research question in Experiment 3 was: *Do the presence of a privacy policy link or the mentioning of cookies influence website visitors' consent decisions?*

Notices in Experiment 3

The base notice for this experiment was the *Category–Non-nudging* notice from Experiment 2 because of the GDPR's "data protection by default" imperative and the ability to obtain consent for specific purposes via checkboxes. We chose a category-based notice over a vendor-based one due to the results of Experiment 2 (see Section 4.5.3). The notice text for this experiment was: "This website [uses cookies | collects your data] to analyze your usage of this site, to embed videos and social media, and to personalize the ads you see. Please select for which purposes we are allowed to use your data. [You can find more information in our privacy policy]." We tested the following conditions:

- *Technical–PP Link:* The original Categories–Non-nudging notice from Experiment 2. It uses both technical language ("uses cookies") and a sentence with a link to the website's privacy policy.
- *Technical–No PP Link:* Same as Technical–PP Link, but the privacy policy sentence was replaced with whitespace to keep the size of the notices consistent.
- *Non-Technical–PP Link:* Same as Technical–PP Link, but using non-technical language ("your data" instead of "cookies").
- *Non-Technical–No PP Link:* Same as Non-Technical–PP Link, but with the privacy policy sentence replaced with whitespace.

For participants who saw a notice with non-technical language, we replaced other occurrences of the term "cookie" in our setup: In the study invitation, "cookie notice" was replaced with "privacy notice," and we adjusted the wording of some survey questions and response options accordingly, as described in Appendix A.1.

4.4.5 *Research Ethics*

Our study was conducted on a website with real visitors, which raises ethical concerns as we did not ask for consent prior to measuring their interactions with consent notices. We did so to ensure ecological validity and be able to capture non-biased results as we expected the majority of visitors to not pay attention to a study consent notice asking them to opt in, which was supported by our findings. Once the data collection for our study was completed, we reactivated our partner website's original consent notice, which did not consider consent decisions made during the time of the study. Hence, recurring visitors who had first accessed the website during one of the experiments would be prompted for consent again.

While our institution did not require IRB review for this study, we ensured that we did not deceive or harm website visitors and their privacy. All displayed consent notices functioned as described and respected the visitor's choice. To

test the effect of no-option consent notices, we had to offer fewer choices than we believe is required by the GDPR. We added a paragraph describing our study to the website's privacy policy. The data we collected was pseudonymized. Logs were stored on the website's server and access was limited to two researchers conducting the analysis and the website's owner. After the study, the data was removed from the server and copied to the researchers' data center.

All visitors were informed about the study after 30 seconds when we showed a notice asking them for participation in the survey. Survey participants were asked for explicit consent and to confirm they were over 18 and wanted to participate. Email addresses of participants who opted to participate in the prize draw were stored separately from the dataset, without the participant ID.

4.4.6 *Data Analysis*

The data we collected comprised website visitors' interaction with our consent notices and survey responses.

EVENT LOGS When we started with data analysis, we noticed inconsistencies in some entries. The event logs created by our plugin indicated that some website visitors had seen multiple notice versions. This could have happened because users had deactivated cookies completely, visited the website in multiple sessions using private browsing mode, or opened the website in multiple tabs simultaneously. For another set of users, we detected multiple screen resolutions, mostly because the screen orientation had changed. Rotating the screen could lead to the notice covering different parts of the website, so we removed these participants to preserve consistency. In total, we removed 2,1 % of participants across all experiments.

SURVEY We considered a survey response complete if the participant had at least answered Q1–Q6 but did not provide a free-text answer to Q7 and Q8. Due to a low survey response rate we received few responses for some conditions. Therefore, we refrained from a quantitative analysis of survey responses but report their counts in Appendix A.1.

In Section 4.5.5, we evaluate responses to the open-ended questions (parts of Q1; Q6–Q8). We coded these responses using emergent thematic coding. Two of the authors independently devised a set of codes for each question and coded the responses. The results were discussed and yielded a final codebook, which was used to re-code all responses. Any remaining disagreements were reconciled by the two coders through discussion. We report the codes and their distribution in Appendix A.1, along with the answers to all closed-ended survey questions.

4.5 RESULTS

Our experiments show that notice position and interaction options, especially paired with nudging, have significant effect on visitors' interaction with the notice, while the effects of (non-)technical language and the presence of a privacy policy link are less prominent. Qualitative survey data reveals some under-

standing of cookies and consent notices, but also widespread misconceptions and the assumption that websites protect visitor privacy by default.

We first describe our participant sample, followed by the results of website visitors' measured interactions with our consent notices, and conclude this section with the results of the survey.

4.5.1 *Data Set and Website Visitors*

Measured interactions

Our cleaned data set contained event logs of 82,890 unique website visitors: 14,135 in Experiment 1, 36,530 in Experiment 2, and 32,225 in Experiment 3. 21.72 % of all visitors accessed the website on a desktop or laptop computer and 78.28 % with a mobile device (of which 5.1 % were tablets)[5]. Overall, 6.95 % of participants used an ad blocker. The rate was much higher on desktop (29.1 %) than on mobile devices (0.8 %). These numbers are consistent with a 2017 report for Germany, which has the highest rate of ad block users in Western Europe (20 % on average), and North America (18 % on average) [207]. For 16.45 % of visitors, we could not detect whether they used an ad blocker. These visitors did not stay long enough on the website for ad blocker detection to complete.

On average, users spent a short time on the website. Pre-study Google Analytics data provided by the partner website showed that 84.81 % of visitors spend less than 10 seconds on the site, 5.21 % 11 to 60 seconds, and 5.83 % up to 3 minutes. Our data set includes all users for whom the event logs indicated a fully loaded site, regardless of how long they stayed on the page, resulting in a high number of "no action" visitors. As described below in Section 4.5.3, the median time until an interaction with any version of the notice was 4 to 8 seconds. About 11,800 users stayed on the page for 10 seconds or more.

Survey participants

The link to our survey was clicked 804 times (168 in Experiment 1, 445 in Experiment 2, and 191 in Experiment 3). We received a total of 110 responses (16 in Experiment 1, 60 in Experiment 2, and 34 in Experiment 3), which means that 0.37 % of the 29,712 visitors who interacted with the notice or stayed on the site for longer than 30 seconds participated in the survey. To get an impression of visitors' expectations about the website's data collection practices, we asked *Q2: What do you think – what data does [the website] collect about you when you access the website?* This question was answered by all participants. Across all three studies, the data most commonly expected to be collected were links clicked on the site (78 %), IP address (65 %), posts read on the site (61 %), and the device used (59 %). Less often mentioned were other visited sites (29 %) and the visitor's place of residence (25 %). 13 % thought the website collected their name, even though the site never asked for it. Only 5 % thought the site did not collect any data about them. These answers indicate that the survey participants had a good understanding of what data websites can collect even without user accounts.

5 We count as "desktop computer" actual desktop machines as well as laptops. "Mobile" devices include smartphones and tablets; the latter were used by 5.1 % of visitors.

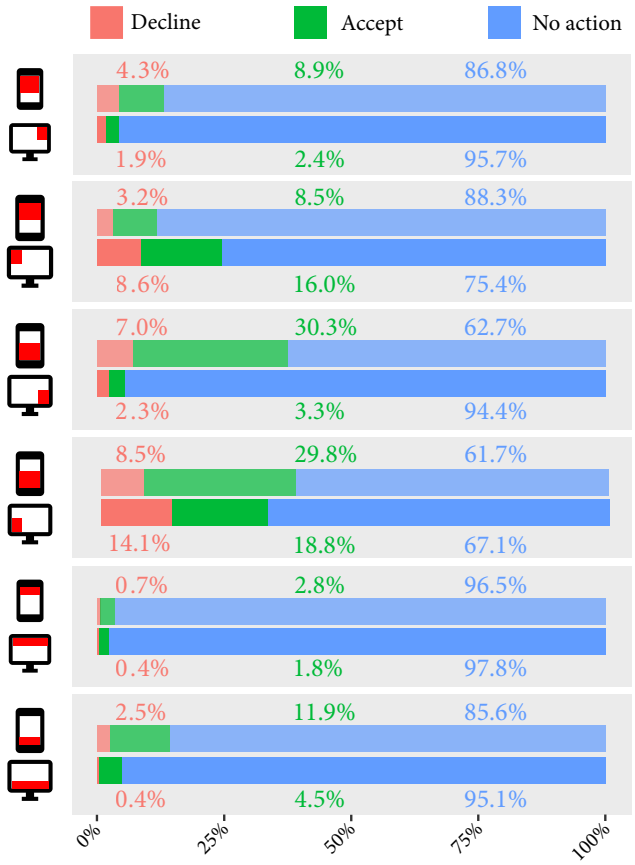


Figure 4.3: Interaction rates in Experiment 1 (notice position), arranged pairwise for mobile and desktop users.

4.5.2 *Experiment 1: Position*

In Experiment 1, we studied the influence of a notice's position within the browser's viewport (see Figure 4.2). We measured the interactions of 14,135 website visitors, resulting in an average of 2,356 people per condition.

4.5.2.1 *Interaction rates*

Highest interaction: bottom left

Figure 4.3 shows how visitors interacted with the consent notices displayed at different positions. Overall, the notices shown at the bottom-left position yielded the highest interaction rates – 37.1 % of visitors interacted with them regardless of device type or choice made. The notice positions most commonly observed in practice, small bars at the top or bottom, resulted in low interaction (2.9 % and 9.6 %, respectively).

Influence of technical parameters

While we were mainly interested in position in Experiment 1, we also analyzed the influence of other variables, such as ad blocker use, screen resolution, browser, operating system, and device type (desktop/mobile). We estimated the effect size of different properties by calculating Cramérs V (CV), and over all visitors the banner position showed the largest effect size (CV = .31). Unless noted otherwise, χ^2-tests for effects in this experiment are statistically significant ($p < .001$).

Ad blocker use also had a small impact on whether someone interacted with the notice. While on average 15.8 % of visitors without an ad blocker interacted with any notice, only 12.6 % of ad blocker users did so, but the effect size was rather small (CV = .11). The impact of screen resolution was much higher on desktop (CV = 0.33) than on mobile devices (CV = 0.16): Only 5.5 % of visitors with screen resolutions of 1,920 by 1,080 pixels or higher interacted with the notice, while the average was 25.6 % for smaller screens. Although the accept–decline ratio varied between conditions, we could not identify a single factor to explain the differences. Across all conditions the number of users who accepted cookies was higher than the number of those who declined.

4.5.2.2 *Discussion*

A possible explanation for higher interaction rates with notices displayed at the bottom is that these notices are more likely to cover the main content of the website, while notices shown at the top mostly hide design elements like the website header or logo. If one uses their thumb to navigate websites on a smartphone, it is also easier to tap elements on the bottom part of the screen than those at the top. An explanation for higher interaction rates with notices displayed on the left of the viewport might be the left-to-right directionality of Latin script: Line breaks cause the information density of a text to be skewed to the left, so consent notices positioned on the left are more likely to obstruct visitors' reading and trigger an interaction with the notice.

We looked for qualitative feedback in the survey responses. In Experiment 1, we received 16 responses, with eight participants having interacted with the notice and another eight that did not. All six participants who answered they had clicked the notice "because it prevented them from reading the website content" had seen a notice shown at the bottom or left side.

Both on desktop and mobile, the notice positioned in the bottom-left corner received the most attention. Thus, we decided to display the notices in Experiments 2 and 3 in the bottom-left corner.

4.5.3 *Experiment 2: Number of Choices and Nudging*

In Experiment 2 we evaluated different options for user interaction and the influence of nudging. There were 36,530 participants in total, and each of the nine conditions was shown to 4,059 website visitors on average.

4.5.3.1 *Interaction rates*

Figure 4.4 provides an overview of the recorded visitor interactions. Compared to Experiment 1, the overall percentage of visitors who interacted with the notice increased (13,8 %–55,3 %), especially on mobile devices, likely because we had increased the font size, resulting in larger notices. The highest interaction rate (55 %) was measured for binary notices on mobile devices.

The experiment revealed a strong impact of nudges and pre-selections. Overall the effect size between nudging (as a binary factor) and choice was CV = .50. For example, even for confirmation-only notices, more users clicked "Accept" in the nudge condition, where this button was highlighted (50.8 % mobile,

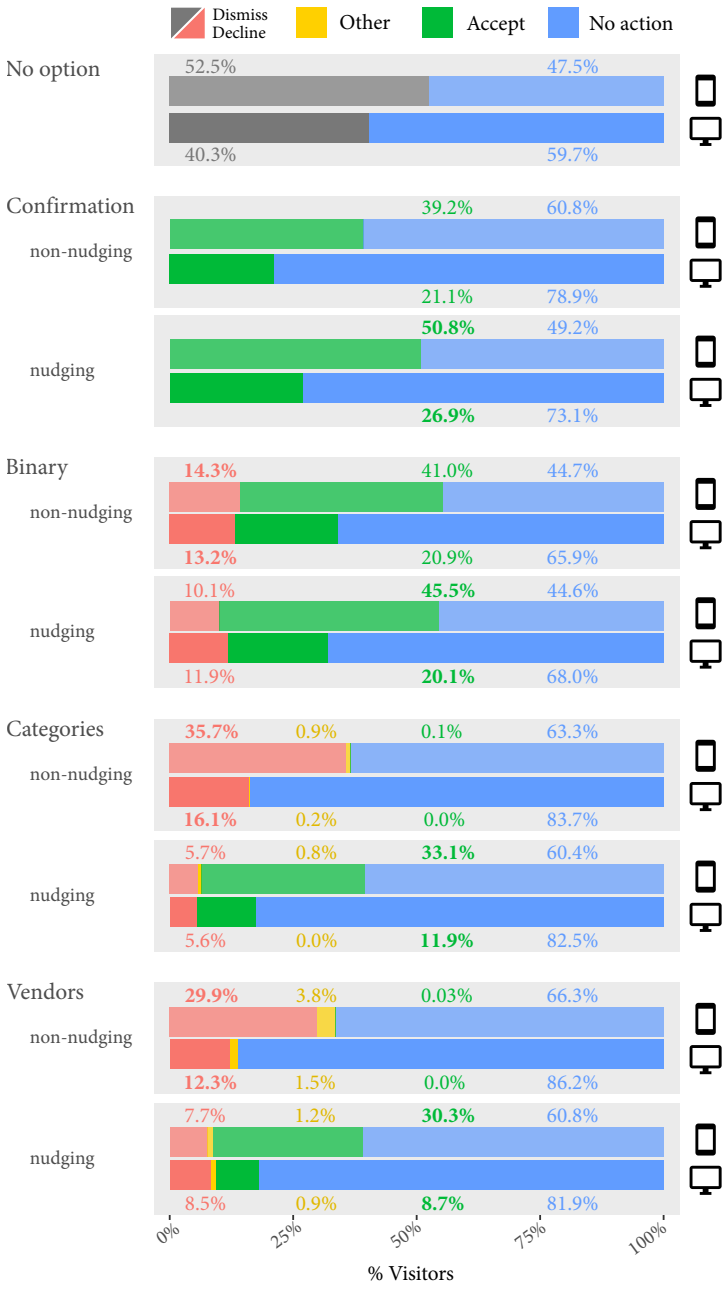


Figure 4.4: Visitors' consent decisions in Experiment 2. "Accept" / "Decline" indicate that (all) options were accepted or declined. "Other" includes those who accepted / declined only some options. Bold figures indicate default options.

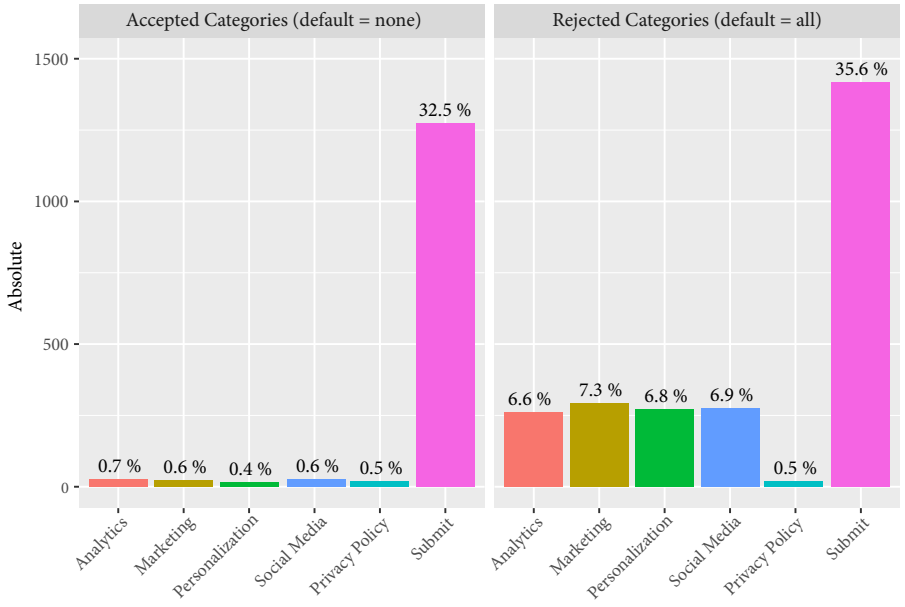


Figure 4.5: Active decisions to allow or decline specific categories in the *Categories–Non-nudging* (left) and *Categories–Nudging* (right) conditions of Experiment 2.

26.9 % desktop), than in the non-nudging condition, in which "Accept" was displayed as a text link (39.2 % mobile, 21.1 % desktop). The effect was most pronounced for category- and vendor-based notices, in which all checkboxes were pre-selected in the nudging conditions, but not in the privacy-by-default conditions. The pre-selected versions led around 30 % of mobile users and 10 % of desktop users to accept all third parties. In contrast, only a small fraction (< 0.1 %) allowed all third parties when given the opt-in choice and 1 to 4 % allowed one or more third parties ("Other" in Figure 4.4), indicating that some users still engaged with the offered choices. No desktop visitors allowed all categories. Interestingly, the number of non-interacting users was highest on average for the vendor-based conditions, although they took up the largest amount of screen space due to six options being offered. We discuss qualitative survey feedback on the category- and vendor-based notices in Section 4.5.5.

4.5.3.2 *Choices*

Engagement by choice mechanism

Results varied in terms of the consent choices visitors made when presented with options (which was the case for all but the no-option and confirmation conditions). Surprisingly, more participants accepted cookies in both binary conditions, where they had the option to decline cookies, than in the non-nudging confirmation condition, where they could only accept cookies or not interact with the notice.

Category or vendor selection

Figures 4.5 and 4.6 show the specific active choices participants made on category- and vendor-based notices, respectively. In the non-nudging conditions, few visitors agreed to data collection through specific categories or vendors if they were not pre-selected. Interestingly, more visitors actively selected specific vendors than categories. Vendors YouTube and Ionic were most frequently selected, even though survey responses (Q6) indicated that Ionic

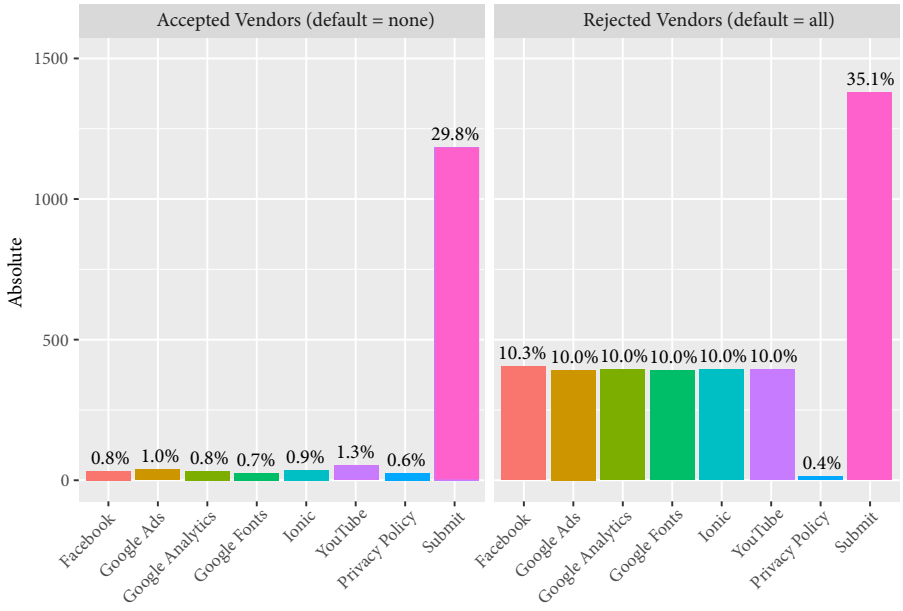


Figure 4.6: Active decisions to allow or decline specific third-party vendors in the *Vendors–Non-Nudging* (left) and *Vendors–Nudging* (right) conditions of Experiment 2.

was not as well known as the other listed third-party vendors. We observe a similar pattern for the de-selection of specific categories and vendors: More visitors unchecked one or more vendors (10.0 %) than categories (6.9 %).

Changes in selection

6 % of visitors who saw a category- or vendor-based notice clicked at least one of the checkboxes more than once. 48 visitors (0,08 %) toggled an even number of times, reversing previous decisions. Interestingly, 47 of those users saw a "nudging" notice, which means that they actively reactivated one of the categories.

Timing data

We also recorded how long it took visitors to submit their choice. Table 4.2 shows timing statistics by choice type. The median time to submit for no-option, confirmation and binary-choice notices was 4–5 seconds; 7–8 seconds for category- or vendor-based notices.[6]

4.5.3.3 *External Validation*

To verify the generalizability of our results, which are only based on visitors to our partner website, we compared our results to internal data supplied by Cookiebot [47], a vendor of a third-party cookie consent solution similar to our category-based notices and also featured in our analysis of consent libraries in Chapter 3. Their data set from February 2019 contained 3 million user logs for 2,000 different websites. The Cookiebot notices also showed purpose categories, so we compared their data to our results for the category-type notices. In their case, some of the checkbox selections cannot be changed by users, as website owners can argue that the use of certain cookie categories is based on different

6 We report the median as the data showed a high standard deviation since we had no way to check when the interaction with a notice started, and sometimes the selection was submitted minutes after the page had been loaded.

Table 4.2: Average time in seconds until users submitted their decision in Experiment 2, if the decision was made within the first three minutes.

Banner	Type	# Users	Mean	Median	SD
No option	n/a	4,174	6.51	4	15.55
Confirmation	non-nudging	2,984	10.65	5	51.25
	nudging	3,634	9.11	4	37.78
Binary	non-nudging	4,134	15.36	4	72.47
	nudging	4,097	13.51	4	75.59
Category	non-nudging	2,523	17.93	8	87.16
	nudging	2,798	13.98	7	64.01
Vendor	non-nudging	2,346	13.76	8	42.38
	nudging	2,741	21.18	7	115.26

Table 4.3: External validation of users' selections in category-based notices (Experiment 2) with data from Cookiebot [47].

Data set	Decision	None pre-selected	All pre-selected
Cookiebot		(n = 1,135,090)	(n = 1,988,681)
	Accept	5.59 %	98.84 %
	Decline	94.41 %	1.16 %
Our Data		(n = 1,239)	(n = 1,380)
	Accept	0.16 %	83.55 %
	Decline	99.84 %	16.45 %

legal grounds (e. g., "legitmate interest", Article 6(1)(f) GDPR). Therefore, (de)selecting all consent-based cookie categories in Cookiebot notices sometimes required fewer clicks, and we were not able to compare decisions we labeled as "other." As shown in Table 4.3, Cookiebot had a slightly higher acceptance rate (5.59 % compared to 0.16 % in our data set) and a lower decline rate when all boxes were pre-selected (1.16 % compared to 16.45 % in our data set). This means that our findings are generally comparable, but specific results may differ based on website and category, which is what we would expect given that privacy preferences are highly contextual [4]. A related 2017 study (n = 300) found that about 3 % of users are willing to accept marketing cookies [230], which falls between the acceptance rates for marketing cookies in our non-nudging (0.6 %) and nudging (7.3 %) conditions.

4.5.3.4 *Discussion*

The results of Experiment 2 show that nudges and pre-selection had a high impact on users' consent decisions. It also underlines that the GDPR's "data protection by default" principle, if properly enforced, could ensure that consent notices collect explicit consent. We further find that even when more choices are offered, most visitors make binary decisions by either agreeing to all or no options. Only very few visitors selected specific categories or vendors, while

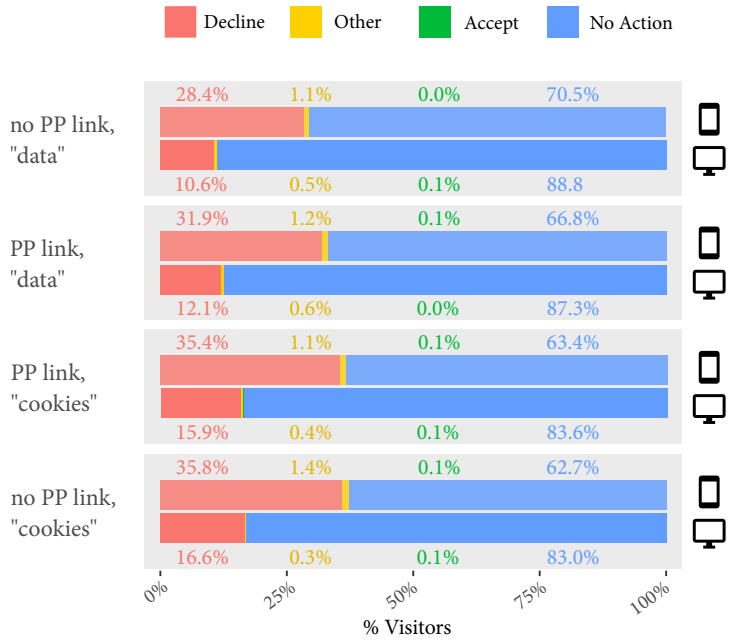


Figure 4.7: Visitors' interactions with the consent notices in Experiment 3. Notices contained technical language ("cookies") or non-technical language ("data") and a link to the privacy policy (or not).

even in the *Binary–Non-nudging* condition a considerable number accepted the use of cookies. An explanation for this behavior might be that people who are somewhat OK with cookie use are not willing to expend effort on enabling them. Another explanation, suggested by previous work [172], is that showing the actual practices decreases the trust in a website and, therefore, leads to more users making an informed decision to decline cookies.

4.5.4 *Experiment 3: (Non-)Technical Language and Privacy Policy Link*

In Experiment 3, we tested four conditions with combinations of (a) the notice including a link to the privacy policy (or not) and (b) the text either referring to "cookies" or "your data" more generally. All conditions were variants of the category-based, non-nudging notice from Experiment 2. Figure 4.7 summarizes the results. The total number of unique visitors in Experiment 3 was 32,225, and each condition was shown to 8,056 visitors on average.

4.5.4.1 *Interaction Rates*

Again, interaction rates were higher for mobile visitors. As in Experiment 2, very few visitors accepted all categories (0–0.1 %), but some (0.3–1.4 %) explicitly allowed one or more category. More people made a choice when technical language was used, i. e., "cookie" was mentioned in the notice. While this difference is significant (χ^2 test, ($p < .01$), the effect size is low (CV = .08), as are the differences between conditions. Presence of the privacy policy link had no significant effect ($p < .08$).

4.5.4.2 *Discussion*

Experiment 3 showed that the mentioning of cookies has a minor influence on users' consent behavior. However, differences between conditions are small. This is not surprising given that most users either submit the default choice or do not interact with the notice at all. We could not confirm previous studies [172] that showed a negative effect on trust in a website when a privacy policy was mentioned, but we found that more visitors decline the use of cookies if a privacy policy is linked. Our findings indicate that position and choice have a more pronounced effect on consent behavior than notice language or pointers to more privacy information.

4.5.5 *Survey Results*

In this section we report the results of the survey (see Appendix A.1) we had invited website visitors to take. Across all experiments, we obtained 110 complete survey responses.

4.5.5.1 *Reasons for (Non-)Interaction with Notices*

In the survey, we asked participants why they had clicked the consent notice or not (Q1). Participants could select multiple reasons.

Reasons for interaction

44 of 61 survey participants who had clicked the notice reported they had done so because they were annoyed by it. 16 thought the website would not work otherwise, and 13 stated they had clicked the notice out of habit. 11 participants interacted with the notice to protect their privacy, 6 for security reasons, and 5 to see fewer ads.

Reasons for non-interaction

49 participants had not interacted with the consent notice, 20 of which reported they had not seen it. Nine thought clicking the notice would not have any effect, six did not care what cookies the website used or what data it collected, and three thought it did not offer enough choices. Two reported not to know what cookies were or what data the question was referring to. 13 participants selected "other" and provided a free-text response. Recurring themes in these responses included that the notices were "annoying [...], so I just ignore them out of frustration" (Participant 2-94[7]) and that participants thought no cookies would be set if they did not interact with the notice. One participant mentioned that they "[found] all of the partners suspicious" (2-255). One had opened the website in a background browser tab, so they had only seen the invitation to take the survey, and two participants reported that the notice had been auto-replaced before they could click it.

4.5.5.2 *Perception of Complex Consent Notices*

We asked survey participants who had seen a category- or vendor-based notice to elaborate on how easy or hard they had found it to make a selection (Q6), in order to learn how they perceived purpose-specific consent mechanisms as required by the GDPR. We received 38 responses across Experiments 2 and 3.

7 The first digit in our participant identifiers denotes the experiment and the second the response ID assigned by LimeSurvey.

Appendix A.1 lists the codes and their distribution for this and the following open-response questions.

A recurring theme in the responses was transparency, as mentioned by 5 participants who had seen a category-based notice: "[I liked] that I could directly select the options without going to the settings. It would be great if this was the default" (3-171), "What I like [here] is that only [the ...] necessary option is selected and all of the others are deactivated" (3-88). One participant with a vendor-based notice stated: "Having options makes me feel secure" (2-619).

However, participants had diverging opinions regarding the notices' clarity. Some found the categories "self-explanatory" (3-118). Others pointed out that "Necessary [from a technical perspective] does not say much. Cookies aren't necessary to view a website" (3-215) and that "something could be hidden" (2-557) behind the Necessary category. 6 (of 7) participants who saw a vendor-based notice in Experiment 2 reported it had "too much text, too many options. I'm interested in the website's content, not in the consent notice" (2-116), and one suggested "it would be perfect to have a button to (de)activate all cookies" (2-199). Seven participants based their choices on privacy considerations: "I don't tick anything. I only need advice [from] the website" (3-108), "I don't want personalized web pages, ads, [... and] pointers to social media" (3-165).

These responses indicate that more complex notices are not necessarily problematic, as long as options are not pre-selected. While some participants expressed concerns, did not trust the categorizations, or found the choices too complex, others appreciated the privacy-by-default approach.

4.5.5.3 *Understanding of Consent Notice Behavior*

The survey further investigated participants' general understanding of how consent notices work and what it meant to accept or decline cookies. This section was identical in all three experiments. The participant was shown the binary notice depicted in Figure 4.1 (a) (bb). Then we asked the following two open-ended questions: *Q7: What do you think happens when you click "Decline"? Q8: What do you think happens when you click "Accept"?*

DECLINING COOKIES For Q7 (Decline), we received 94 responses across the three experiments. We identified ten themes. The most prominent expectation was that declining cookies would prevent access to the website (28 responses): "I don't get access to the desired information" (1-282), "The site closes itself and you are redirected to the search engine" (2-685). 17 other participants expected parts of the website not to work: "I won't be able to use some functionality because [...] cookies fund the website" (2-255). Only 4 participants explicitly mentioned that they would be able to access the site, stating, for example, "Normally I can continue to navigate the site. It has only happened twice that [a] site has kicked me out. But online shopping [is] difficult if you don't agree" (2-94). 3 participants expected no collection or processing of personal data to take place when cookies are declined but still had doubts: "I hope that no data is collected" (1-177, 1-121, 3-216). 12 expected the site to behave as if "Accept" had been clicked: "I guess my data is still collected" (1-170), "Nothing, of course. Me not accepting cookies does not mean that the site uses fewer or no cookies or does not collect any data about me" (2-630). Other recurring

themes in the responses include the expectation to see fewer ads, a focus on the technical aspects ("no cookies are evaluated" [3-217]), and if the notice would dis- or reappear. Appendix A.1 contains details about the distribution of codes across experiments.

ACCEPTING COOKIES For Q8 (Accept), which was also answered by 94 participants (not all the same respondents as for Q7), we also identified 10 themes. 29 participants expected that their personal data would be collected and/or processed: "my behavior on the website is stored and analyzed" (2-216), "my data is shared with who knows what third parties [...] Facebook, Google, marketing / market research / ad analytics [...]" (2-557). 19 responses focused on technical aspects: "a cookie is set which recognizes me when I revisit the website" (1-250). 21 participants stated the website would only work if they allowed cookies: "I can read the article" (2-53), "I can continue to use the website" (2-405). Other themes included effects on the consent notice only ("the banner disappears" [2-675]), personal data being collected for advertising, user profiling, and other purposes, for example, "sale to third parties" (3-171), "influencing Internet algorithms" (1-269), and "any purpose" (1-207, 3-64). 7 participants believed it made no difference what was clicked but did not specify what that "default" behavior of the website would be.

Misconception

These answers indicate that our participants had some understanding of how cookies are used – for example, to recognize recurring visitors and for ad tracking and targeting. Concerningly, almost a quarter of participants thought they had to accept cookies before they could access a website, which may be due to negative experiences on some sites influencing general expectations and behavior across websites. A transparent and GDPR-compliant consent notice should inform users which website functionality may not work as intended if cookies are declined.

4.6 DISCUSSION AND LIMITATIONS

We conducted three experiments evaluating the effects of cookie consent notices' position, choices, and content on people's consent behavior. In the following we describe recommendations based on our findings and discuss limitations of our approach.

4.6.1 *Recommendations*

Design of consent notices

Our experiments investigated different notice positions, offered choices, and aspects of the wording of cookie consent notices. Future guidelines for consent notices should consider the following recommendations:

POSITION Experiment 1 showed that the position of a notice has a substantial impact on whether a website visitor engages with the notice. A dialog box in the lower left corner (on desktop) or the lower part of the screen (on mobile) significantly increases the chance that a user makes a consent decision. While we had expected higher interaction rates on mobile devices for this position since it is easy to reach with the thumb, we were surprised by the impact on

desktop users, given the general wisdom that content in the top left receives the most attention in cultures with left-to-right writing. This result could be related to our partner website, like many websites, displaying a header, which shifted content to lower parts of the screen. This experiment further shows that the second most common notice position observed in practice, the top position (see Table 4.1), results in notices being ignored by users.

CHOICES Our results from Experiment 2 showed that nudging (highlighting "Accept" buttons or pre-selecting checkboxes) substantially affects people's acceptance of cookies, providing clear evidence for the interference of such dark patterns with people's consent decisions. Given a binary choice, more visitors accepted cookies than declined them, which could be evidence for the adverse effects of consent bundling on consent decisions, which is not allowed under the GDPR. Surprisingly, rejection rates in the vendor- and cookie-based conditions were close to those in the binary condition, although visitors had to make five to six additional clicks to reach the same goal. This suggests that people who want to decline cookies are willing to expend extra effort.

Moreover, the survey answers show that participants think that no data is collected *unless* they make a decision, showing that privacy by default is the expected functionality, although this is not the current practice.

TEXT While we did not see an effect in Experiment 3 from including a privacy policy link in the notice, we found that mentioning "cookies" made more users reject the data collection. The negative effect of mentioning cookies can very well be related to the fact that Internet users have in general a negative feeling about them [106, 147].

Unsatisfactory state of consent in the wild

It is clear that the current ecosystem of mechanisms for websites to prompt for user consent – with a plethora of combinations regarding the provided information, the granularity of options, and how and if visitors' choice is enforced – provides no real improvement for visitors' data privacy compared to pre-GDPR times. At the same time many aspects of how to obtain consent under the GDPR and ePrivacy Directive are still unclear, with regulators publishing differing guidelines on how to obtain consent, the online advertising industry developing and updating proposals for consent frameworks, and legal and technical scholars evaluating them. While some claim [231] that many core mechanisms of the online advertising industry are not compatible with the GDPR at all, the regulation so far has only partially affected how companies process personal data [279]. We hope that our results can inform future discussions, not only with recommendations for the design of consent notices. Given that at the moment few website visitors are willing to give consent to any form of processing of their personal data, we think that the business model of online behavioral advertising, which targets ads based on large amounts of personal data, should be challenged and alternative models like privacy-friendly contextual advertising [65] or other ways of monetization for web services need to be promoted and developed.

4.6.2 *Limitations*

Bias in sample

Our study has some potential limitations. First, our sample is biased, as we conducted all experiments on a German-language e-commerce website whose visitors may not be representative of the general public. However, our partnership with this website gave us control over the notice implementation and access to a high number of unique visitors. We validated some of our results with data from Cookiebot which showed similar results (see Section 4.5.3.3). Overall, it seems that the participants in our sample are more inclined towards rejecting cookies. We have to assume that in general a higher percentage of website visitors may allow cookies. Our field study did not allow us to collect more detailed information about visitors, such as their specific device, the size of the notice on the screen, or how long they stayed on the website, which could potentially have an effect on consent behavior.

Website dwell time and lack of user accounts

Furthermore, many visitors did not interact with the notice at all and spent only a short period of time on the site. While this could be related to the notice, it is not unusual that most visitors leave a site after a few seconds. Liu et al. [164] showed that website dwell time has a negative aging effect. Users first skim a site to decide whether they will stay on it. Since we were not able to measure the exact time visitors stayed on the site, we included all users for whom the logged data indicated a fully loaded page, which results in a high number of "no action" visitors. From a legal perspective the time spent on the site does not affect the need to request consent. Our partner website also does not have user accounts. Past research has shown that visitors tend to underestimate the amount of personal data collected by websites on which they do not create an account and enter personal data [222]. This may cause them to underestimate the privacy implications of allowing cookie use, but we did not see evidence for this in the survey responses.

Self-selection bias

Responses to our voluntary survey are likely biased due to participants' self-selection. Responses to the question about possible data collection suggest that participants had a good understanding of the technical background or an interest in privacy. Of the survey participants, 61 had previously interacted with our consent notices and 49 had not, showing that the results are only partially biased towards those who care about notices. We considered this bias when interpreting results.

4.7 CONCLUSION

We conducted the first large-scale field study on the effect of cookie consent notices on people's consent behavior. Cookie notices have seen widespread adoption since the EU's General Data Protection Regulation went into effect in May 2018. Our findings show that a substantial amount of website visitors are willing to engage with consent notices, especially those who want to opt out or do not want to opt in to cookie use. At the same time, position, offered choices, and nudging substantially affect people's consent behavior. Unfortunately, as shown in Chapter 3, many current cookie notice implementations do not make use of the available design space, offering no meaningful choice to consumers. Our results further indicate that the GDPR's principles of data protection by

default and specific, purpose-based consent would require websites to use consent notices that would actually lead to less than 0.1 % of visitors actively consenting to the use of third-party cookies.

Part III

WEBSITES' PERSPECTIVE: PRACTICES IN THIRD-PARTY USE

5

CONSIDERATIONS IN THIRD-PARTY ADOPTION BY WEBSITES

5.1 INTRODUCTION

Privacy-by-design to overcome the need for consent notices

With website visitors becoming increasingly annoyed by ubiquitous, complex, and intransparent consent notices, websites could choose another path to reconcile user experience with legal compliance: follow the GDPR's "data protection by design and by default" principle (Section 2.2.1.5) and adopt solutions for more privacy-friendly integration of the desired website functionality (see Section 2.1.4) that do not collect as much information from visitors and, thus, may not require user consent. The observed lack of change in third-party use on websites under the GDPR raises the question to what degree websites already make use of such technology or at least consider it when deciding how to integrate a desired website functionality, and what could be done to foster use of privacy-friendly alternatives to popular third-party services.

Role of the human factor in third-party adoption

Web developers – and people in similar roles related to a website's creation and administration – are a crucial part of the third-party tracking ecosystem, as it is them who integrate third parties into websites, thus enabling them to track visitors' behavior across the Web. Hence, to foster the adoption of privacy-by-design in the context of third-party services on websites, it is important to understand to what extent people tasked with the creation and maintenance of websites are *aware* of the privacy risks of third-party use and if they consider visitors' privacy both in the decision that leads to the *selection* of third-party services and in *integration* itself.

Existing focus on the mobile domain

Though prior work has studied the history [155, 288] and prevalence [63] of third-party web tracking, little is known about the decision processes behind the use of third-party services on websites and if website visitors' privacy is considered in the process. Previous work that has studied developer behavior in adopting [215] and updating [232, 233] third-party libraries has focused on smartphone apps, for example, investigating developers' privacy considerations in their use of mobile advertising networks [183, 266], their awareness of data collection through third-party tools for unspecified types of functionality including ads and analytics [15], and their adoption of alternative APIs that preserve location privacy [136]. Third-party services and libraries for websites differ from those for the mobile ecosystem in their availability for a greater variety of purposes, the potential for higher technical complexity, and higher sophistication of advertising ecosystems [129, 156, 280]. Websites also lack apps' distribution through a centralized platform, whose requirements may shape developers' understanding of privacy aspects, including what data is considered sensitive [269]. On the Web, the omnipresence of consent notices that implement IAB Europe's Transparency and Consent Framework (TCF) [114] and often list a site's third-party vendors could have led to higher awareness of data collection through third parties on websites compared to the mobile space, where consent prompts are much less prevalent [144].

Our contribution: Online study of privacy considerations in the use of third-party services

In this chapter, we address this research gap with findings from a mixed-methods online study with 395 participants who were involved in the design, development, deployment, maintenance, or management of websites. We combine survey answers with web privacy measurements and investigate how ten website functionalities associated with frequent use of third-party services have been integrated into websites and how visitors' privacy was considered in the process. We go beyond prior work by exploring privacy considerations between different types of functionality that may not be equally prone to third-party use [162], as well as factors that influence the adoption of first- vs. third-party solutions to integrate a functionality.

More specifically, we make the following contributions:

- We extend web privacy research on the prevalence of third-party services by contrasting their use with first-party integrations for different purposes, regarding their prevalence, factors that drive use of first vs. third-party solutions, and consideration of alternatives. We find that the decision in favor of third-party services, as in the mobile domain [233], is driven by ease of integration, features, cost, and familiarity with a service, while privacy rarely is a decisive factor. However, we find use of privacy-friendly integration for web analytics and programming / design resources, and self-hosting tends to be the primarily considered alternative to third-party solutions, rather than another third party.
- Like prior work on cryptographic APIs [1] and mobile ad networks [183], we find that adjustments to a service's default configuration are rarely reported. However, participants who did change defaults often did so in response to privacy-related court rulings or guidelines by data protection authorities.
- We find higher awareness of data collection pertaining to a third-party service's core functionality, such as financial information for payment or behavioral data for analytics, whereas awareness is lacking for data collected in less prominent contexts, particularly the transmission of IP addresses and device information.
- From a methodological perspective we contribute to the ongoing discussion about ethics in security and privacy research by discussing implications and lessons learned from using public GitHub data to recruit people involved with web development, a method previously used by developer-centered research [1, 2, 102, 187, 233, 244, 245, 265, 268, 294].

Our findings demonstrate the need for researchers and the web development community to raise awareness of the privacy risks associated with the use of third-party services on websites, as well as the need for clearer regulatory guidance and requirements for privacy-friendly defaults.

5.2 RELATED WORK

Previous work has studied the prevalence and evolution of third-party web tracking and developers' privacy behaviors in third-party use in the mobile app ecosystem.

5.2.1 *Evolution of Third-Party Web Tracking*

Web tracking has been studied extensively, including the prevalence of third-party tracking services on websites. Tracking has been identified since 1996, and since then increased in prevalence and complexity [155], with the most popular services covering up to 75 % of websites in 2015 [288] and hundreds of different known tracking services [226] whose use increases with website popularity, and visible differences between regions and website types [120]. Large-scale investigations confirmed that more than half of websites leak user data or load third-party scripts [160]. As shown earlier in Chapter 3, the GDPR going into effect in May 2018 primarily was accompanied by an increase in the prevalence of cookie consent notices, while actual tracking practices did not change much. Other changes identified by concurrent work could not be directly attributed to the GDPR [251]. While there were clear differences between website visits from US or European users, implying that companies collect less data from the latter [48], previous research, overall, did not find significant positive developments in third-party web tracking due to the GDPR.

5.2.2 *Developers' Privacy Considerations*

Developers' considerations of users' privacy have been studied in different contexts, but there are few insights into *why* specific third-party services are used in web development. Previous work found that developers of mobile apps are often unaware of third-party data collection [15] and, therefore, tend to collect more data than necessary. Developers also showed a limited perception of privacy threats, often based on their organization's guidelines [108]. Mhaidli et al. investigated how and why mobile app developers use and choose ad networks and whether they consider associated risks for users [183]. They found that developers see advertisements as the only viable way to monetize their apps and consider ad networks to be responsible for protecting app users' privacy, not themselves. Tahaei et al. confirmed this and showed that app developers find existing privacy information and controls confusing and hard to use [266]. Roth et al. [228] investigated developers' struggles with the use of Content Security Policy (CSP) and found that third parties interfering with the deployment of the mechanism (see Section 2.1.3) has prompted a third of the interviewed developers to consider alternatives to the problematic third party, such as self-hosting or upgrading to the latest version of a design framework that no longer requires jQuery. Other studies investigated public forums to learn how developers deal with privacy regulations and changes to them, finding that they mostly try to uphold standards defined by large companies [269] or are focused on recent changes or events [159] when discussing privacy. When asked to solve privacy-focused tasks, developers tend to use better-documented alternatives and copy examples, which could be adopted by privacy-friendly services [136]. They often struggle with embedding privacy into their application due to a lack of knowledge, privacy contradicting app requirements, or task complexity [213, 245]. Another problem are third-party vendors' competing business interests, leading them to employ dark patterns that steer developers towards privacy-unfriendly defaults [267].

We add to this body of knowledge by investigating if and how people involved in the creation or administration of websites consider privacy when selecting and integrating third-party services.

5.3 METHOD

To investigate the privacy practices and decision processes behind third-party use on websites, we conducted a mixed-methods study consisting of an online survey with 395 people involved in the creation and administration of websites, paired with an analysis of participants' websites, if provided in the survey. At the heart of our survey, we asked participants about their selection and implementation considerations for up to three out of ten website functionalities often integrated using third-party services. Next, we describe the set of investigated functionalities, our survey design, recruitment process, research ethics, and data cleaning and analysis.

5.3.1 *Website Functionalities of Interest*

In Section 2.1.5 we already described how we identified common use cases for third-party services from existing categorizations in related work and web tracking projects. To further reduce the number of categories for the purpose of this study, we took the consolidated list (Table 2.1) and removed categories that apply only to a first-party context (e. g., hosting, distribution) or only make sense combined with other categories (e. g., tag management). Our final list comprised ten common website functionalities, for many of which privacy-friendly implementation options exist, as outlined in Section 2.1.5: Advertising, analytics, customer interaction, embedded media, user login / authentication, payment, privacy plugins / forms, programming and design resources, social media integration, and website protection.

5.3.2 *Survey Design*

Our survey was inspired by the work of Mhaidli et al. [183] and consisted of five parts. It was conducted in English and implemented on a self-hosted LimeSurvey instance. To prevent early priming about privacy, we framed the survey as exploring practices in the selection and use of web technologies on websites and only introduced questions about privacy and data collection practices in Part 4. The full survey questionnaire can be found in Appendix A.2.

Survey structure

Figure 5.1 shows the high-level structure and logic of our survey. Part 1 assessed participants' background regarding their work on websites, including their experience with the ten functionalities of interest (Q1-3).

To provide context for the rest of the survey, Part 2 asked participants to think of one specific website they had recently worked on and remembered well and to only keep this website in mind for subsequent questions. Participants could optionally provide the website's URL (Q2-0). The survey consent form explained that this information would be used to check which web technologies

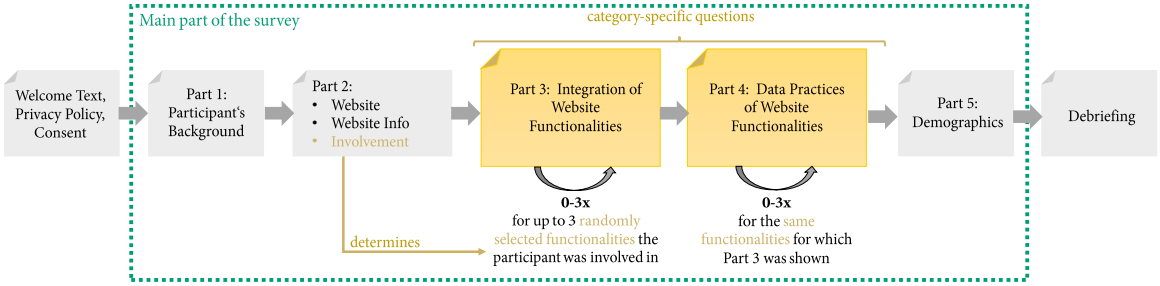


Figure 5.1: Structure of our survey. Any functionality for which the participant had indicated some type of involvement in Q2-7 was placed into a random draw, and for up to three randomly selected functionalities, survey parts 3 and 4 were shown.

were present on the website. At this point we investigated the methodological question if requiring participants to provide a website had an effect on dropout rates: We made Q2-0 mandatory for half of GitHub-recruited participants (see Section 5.3.3) but could not find evidence that this had an impact on dropout rates or the willingness to provide a website.

Part 2 proceeded to ask about website metadata, including the country it was based in, the participant's role with regard to the website, and which of the ten functionalities of interest (see Section 5.3.1) were present on the site (Q2-6). To balance level of detail and survey length, we chose to display more detailed questions only for up to three functionalities. For this, Q2-7 asked, for each functionality indicated to be present in Q2-6, to what degree the participant had been involved in the decision of how this functionality should be integrated (selection), in the integration process itself, and in maintenance or management of the integrated solution. From the functionalities for which any kind of involvement had been indicated, three were randomly selected, for which Parts 3 and 4 of the questionnaire would be shown.

Part 3 investigated how a functionality was integrated in terms of first- vs. third-party solutions and, if applicable, embedding mechanism. It also asked about the underlying decision process including reasons for selection and considered alternatives, information sources, and the people involved.

Part 4 explored participants' understanding of the data collected through third-party services and efforts made to protect visitors' privacy in the integration process.

Finally, Part 5 asked demographic questions and if participants had received training or educated themselves on data protection or privacy. At the end of the survey, participants were debriefed about the study's privacy focus and given the option to either withdraw from the study or to submit their answers. Six participants withdrew at this point.

Survey testing

To assure survey quality, we first conducted "think-aloud" cognitive interviews with seven web developers and two content creators, recruited via convenience sampling. After each interview, we addressed identified issues and repeated this process until no further issues emerged. A pilot launch of the survey with 101 participants recruited from GitHub (see Section 5.3.3) did not yield evidence of any remaining issues, so we proceeded with data collection.

5.3.3 *Recruitment*

Our recruitment approach was guided by the goal to obtain different perspectives on website functionality integration. We leveraged two recruitment channels to obtain a diverse sample: websites' contact information to reach individuals in a range of website-related roles, and GitHub to reach web developers. People were eligible to participate if they were at least 18 years old, worked on websites in some capacity (e. g., website design, development, deployment, maintenance, management), and were comfortable taking the survey in English. Participation was voluntary and uncompensated.

Website-based recruitment

To cover a diverse range of websites in recruitment, we searched the top 100,000 domains on the Tranco list of popular websites[1] [152] for email addresses related to a website's technical administration.

Conducting a meaningful, context-based search for email addresses first required us to identify the languages of the websites in our set. For this, we used the Python library `langdetect` on the homepages of the domains on the Tranco top 100,000 list. For each of the 20 most common languages on this list[2] we drew a random sample of 10 websites with country code TLDs from countries where this language is widely spoken (e. g., `.pt` and `.co.br` for Portuguese), and, using DeepL and Google Translate, identified and collected the names of the links that led to the sites' subpages containing privacy-related policies, terms of service, and contact information. The final, consolidated list comprised 103 terms for privacy-related pages and 72 terms for pages with contact information.

Using this list, we visited each domain on the Tranco top 100,000 list in October 2020 with Open Web Privacy Measurement (OpenWPM) version 0.13 [63], downloaded the corresponding subpages and the homepage, and searched them for email addresses with a regular expression. Since websites often list contacts responsible for the content (e. g., editors on news pages, politicians on government sites) rather than administration, we excluded subpages with more than four email addresses. After removing duplicates, invalid email addresses, and subpages with more than four addresses, we were left with 109,862 unique email addresses for 53,496 websites.

GitHub-based recruitment

Previous work studying web developers' security and privacy practices has used public GitHub repositories to recruit developers on a large scale [1, 2, 102, 187, 233, 244, 245, 265, 268, 294]. We also used this approach because it allowed us to recruit people likely involved with web development without hand-picking them, as would have been the case for one-by-one contact on platforms such as LinkedIn. Though prior work is not always clear on where exactly on GitHub users' email addresses were collected (options include commit email addresses and users' profile pages), from discussions with authors of some previous studies we know that the use of commit email addresses is common. Following

1 List from September 1, 2020 (`https://tranco-list.eu/list/64WX`).

2 By frequency: English, Chinese, German, Russian, Spanish, Korean, Japanese, French, Vietnamese, Persian, Portuguese, Arabic, Italian, Turkish, Indonesian, Polish, Dutch, Danish, Greek, and Catalan. These cover 97.3 % of the websites we could download and whose language could be identified.

this previously used method, we analyzed commits made into public GitHub repositories in August 2020 to identify e-mail addresses of people working on websites, as indicated by the respective commit including file extensions related to web development: `.js`, `.php`, `.css`, `.html`, and `.htm`. Anticipating a low response rate, we sent invitations to 37,000 email addresses, in addition to 12,000 contacted during pilot testing.

5.3.4 *Research Ethics*

Prior data protection review

Prior to conducting the study we looked into opportunities for ethical and data protection review at our institutions. At the time this study was designed, conducted, and evaluated, the involved researchers were affiliated with Leibniz University Hannover (LUH) and Ruhr University Bochum (RUB), both located in Germany, and the University of Michigan (U-M) in the US. RUB only had an Institutional Review Board (IRB) for research in psychology, which was not meant to be mandatorily consulted by security and privacy researchers. LUH's IRB only targeted project proposals, not individual research papers. The researcher from U-M did not directly work with raw response data or interact with participants and confirmed with U-M's IRB that their oversight and approval was therefore not required. Nevertheless, we followed best practices for research conduct and transparency. To ensure GDPR compliance of our study, we consulted RUB's and LUH's data protection officers (DPOs). They both independently considered our study design and specifically the approach for GitHub recruitment to be covered by the GDPR's research privilege.

Naming a website

In Q2-2 we required some participants to provide the URL of a website they had worked on, following Mhaidli et al.'s study design [183]. We explained in the initial consent form that this data would only be used to check the provided website for the presence of third-party services. Participants required to fill this field were able to drop out or proceed without penalty by entering arbitrary input.

Recruitment

Regarding recruitment, we carefully considered the implications of sending email invitations to website contacts and GitHub developers at a large scale. As mentioned above, the two consulted DPOs considered this recruitment approach to be GDPR-compliant. We contacted each email address only once (i. e., we did not send any confirmations or reminders) and gave email recipients a one-click option to opt out of further contact. Still, we received a small number of emails with negative sentiments from people who were not aware that their public GitHub commits contained their email address. Upon this feedback we put up a page on our institution's website that explained our study, why the GitHub-recruited recipient's email address was visible in commits into public repositories, and what steps could be taken to hide it.

Despite these efforts, one recipient filed a complaint with our state's data protection authority (DPA), upon which we immediately stopped recruitment via GitHub, rather than waiting for the outcome. Three months later the DPA informed us that they did not consider the GDPR's research privilege to apply, because GitHub users, who are often unaware of their commit email addresses being publicly available, do not expect to be contacted via these addresses

for the purpose of scientific research. We discuss the concrete problem with GitHub's mechanics for email addresses in more detail in Section 5.5.4. The DPA advised us to refrain from future recruitment via public GitHub commits but did not take formal action.

Ethical "lesson learned"

When we designed and launched the study, ethical concerns with recruitment via public GitHub commits were not obvious: The method was established in the community [1, 2, 102, 187, 244, 245, 294], even post-GDPR [233, 265, 268], and had passed ethical or IRB review at different universities in the US, Europe, Australia, and at the NIST Human Subjects Protection Office. As such we followed established research practice at the time, as well as sought consultation and approval regarding the GDPR from two DPOs from different institutions, who independently concluded the recruitment method to be covered by the GDPR's research privilege. In hindsight, we agree with participants' and the DPA's concerns regarding GitHub recruitment, which is why we decided to fully discuss our experience here. We consider this aspect of our work a valuable lesson learned for the community in how legal or ethical assessment of established study methods can – and should – evolve. We discuss the implications for future work in more detail in Section 5.5.4.

We want to stress that all participants whose data is reported in this chapter provided their information with informed consent, obtained both at the beginning of the survey and at the end after debriefing about the study's privacy focus. The issue pointed out by the DPA lies with the recruitment method, not with the data we received from the willing and consenting survey participants.

5.3.5 *Data Analysis*

Across all recruitment phases, 2,177 people opened the survey link, 667 proceeded past the welcome page, and 452 completed the survey.

Data cleaning

Out of these, we removed 41 that had not seen Parts 3 and 4 due to a lack of reported involvement, nine who selected contradictory levels of involvement, and seven who provided multiple websites. To increase data quality, we examined participants' response times. The average survey completion time was 20:42 minutes. We did not observe any suspicious patterns and thus did not remove any answers. This left us with a total of 395 valid responses.

Two authors inspected all open-response "Other" answers and re-coded answers that matched existing closed-ended options after discussion and mutual agreement. For website analysis, one author inspected all provided URLs (Q2-0) and removed all answers that were not URLs (e. g., "client confidential") or could not be resolved to a website.

Survey responses

Two of the authors applied thematic analysis [36] to the answers to open-ended questions. First they independently reviewed the data to identify recurring themes and created individual codebook drafts for each question. Next, they discussed these drafts and merged them into a first joint codebook. All data was then jointly coded by both researchers, who discussed problematic cases until an agreement was reached, which at times required refining codes' definitions and scopes and, thus, revisiting previously coded answers.

We did not compute inter-rater reliability, as the number of responses was small enough to not require splitting up between multiple coders [181]. Each

open-ended response could be assigned one or more codes, as participants often mentioned more than one relevant talking point.

Website analysis

To assess to which extent participants' responses about websites' integrated functionalities matched actual practice, we analyzed the provided websites with OpenWPM [63]. As the amount of web tracking [48] and consent banners [114, 284] might be dependent on the visitor's location, we performed OpenWPM crawls simultaneously from vantage points subject to the GDPR (Frankfurt, Germany), the CCPA (San Francisco, California, USA), and not subject to either (Bangalore, India).

Previous work has found that only visiting the homepage of a domain does not provide an accurate picture of a site's tracking practices [280]. While there are functionalities that result in the third-party service being accessed from all subpages, such as web analytics or social media buttons in a sidebar, others are only present on specific subpages. Examples include embedded maps on a company's "About" page showing the company's physical locations or contact forms with a CAPTCHA on a "Contact" page. For this reason, we accessed the front page for each provided URL, searched it for links to subpages, and visited up to 100 randomly selected unique pages, as recommended by Urban et al. [280]. For each page, we collected all HTTP(S) requests to determine the third-party services used on the website and used the WhoTracks.me database [139] to categorize and map them to our functionality categories to allow for comparisons between survey responses and website measurements.

Website metadata

Finally, we compiled metadata on the provided websites: top-level domains (TLDs), topic classification, and popularity. For classification we used McAfee's Real-Time Database [176] due to its extensive coverage [283]. Popularity was assessed with the same Tranco ranking we used for website-based recruitment. For websites hosted on platform subdomains (e. g., `xxx.github.io`, `yyy.herokuapp.com`) we considered these as distinct TLDs and counted each platform only once in our popularity statistics.

Statistical methods

For data analysis we mainly rely on descriptive statistics because the variance in response counts per website functionality would cause statistical tests to often be underpowered. Where statistical tests are appropriate and possible we used Fisher's exact tests to check if differences between categories were significant and corrected for multiple tests with the Benjamini-Hochberg procedure.

5.4 RESULTS

Our results show that, as in other domains, user privacy is rarely considered in web development. Yet, we do find influence of regulators' guidelines for some types of functionality, and self-hosting is a prominently considered alternative to third-party use. We also find a widespread lack of awareness that third-party use implies transmission of IP addresses and device information to the third party.

5.4.1 *Sample*

We first describe the sample of 395 participants and 361 websites they provided to support the main part of the survey.

5.4.1.1 *Participant Demographics and Background*

Demographic data

Table 5.1 summarizes participants' demographics (Part 5 of the survey) and background in their work with websites (Part 1 of the survey, Q2-1, and Q2-2). Participants predominantly identified as men (85.1 %; Q5-2), are most frequently in the 18–24 (33.4 %) or 25–34 (30.6 %) age ranges (Q5-1), and the majority holds a bachelor's degree (35.2 %; Q5-3). Most reported degrees (Q5-4) were in technical fields, with the most common non-technical degree being in business/economics (10.4 %). This is consistent with demographic surveys of people working with web technologies, whose large majority are men, typically in the 24–34 age range, holding a bachelor's degree in technical fields [50, 104, 254, 300].

Participants' background

Participants' work with websites (Q1-2) was most frequently conducted in a full-time position (41.8 %), though freelancing and part-time employment were also common, as was non-paid work (hobbyist 31.4 %). In the last three years, participants had mostly worked on 2–5 websites (43.8 %; Q1-1). As for previous experience with the ten website functionalities (Q1-3), all but one participant reported at least one functionality, with a mean of 5.28 (SD 2.37, median 5). Experience with front-end programming or design libraries (83.0 %) and user login or authentication (80.5 %) was most common, while the fewest participants had worked with privacy plugins (29.9 %) and advertising (23.0 %). Participants held on average 3.4 different website-specific roles (SD 2.58, min 1, max 13, median 3; Q2-1) and most often worked as (web) developer, programmer, or software engineer (85.3 %). Other frequently reported roles include administrator / web operator, user experience (UX) design, content creator or contributor, and product or project manager. Most participants worked alone (35.7 %) or in teams of sizes 2–5 (35.7 %) (Q2-2). Prior privacy training (Q5-5) had been received by 42.0 % of participants. The most common resources of such training were self-study (38.6 % of participants with training), employer training, courses at a university or school, and other non-online courses, including certifications such as Certified Information Systems Security Professional (CISSP).

Table 5.1: Participants' demographics (Part 5 of the survey) and background (Part 1 of the survey, Q2-1, and Q2-2). [C] indicates coded open-ended answers, [M] indicates multiple-choice questions or multiply assigned codes for which (response) counts can sum up to more than 100 %. Percentage values are relative to the total number of survey responses (n = 395). For the coded open-ended answers to the type of privacy training received (Q5-5; bottom left, indented list), percentage values are relative to to the number of participants who indicated to have received prior privacy training (n = 166).

	Demographics	n	%
Age	18–24	132	33.4
	25–34	121	30.6
	35–44	76	19.2
	45–54	30	7.6
	55–64	20	5.1
	65–74	5	1.3
	75+	1	0.3
Gender[M]	Woman	40	10.1
	Man	336	85.1
	Nonbinary	4	1.0
	Self-described	3	0.8
Education	No schooling completed	5	1.3
	Some high school, no dipl.	14	3.5
	High school graduate	57	14.4
	Some college, no degree	39	9.9
	Techn. / vocational training	13	3.3
	Associate degree	5	1.3
	Bachelor's degree	139	35.2
	Master's degree	77	19.5
	Professional degree	9	2.3
	Doctoral degree	21	5.3
	Other	4	1.0
Field of Degree[M]	Computer / information sc.	222	56.2
	Mathematics	53	13.4
	Engineering	89	22.5
	Life sciences	19	4.8
	Physical sciences	26	6.6
	Social sciences	23	5.8
	Education	19	4.8
	Law	2	0.5
	Psychology	5	1.3
	Business / economics	41	10.4
	Liberal arts / humanities	23	5.8
	Art / music	10	2.5
	Journalism	7	1.8
	Vocational	3	0.8
	Not applicable	24	6.1
	Other	9	2.3
Privacy Training[C,M]	Yes	166	42.0
	Self-taught	64	38.6
	Employer training	39	23.5
	'Learning by doing'	10	6.0
	University / school	18	10.8
	Online courses	11	6.6
	Other courses	25	15.1
	Professional network	7	4.2
	Other	5	3.0
	No	189	47.8

	Background	n	%
# Websites	1	18	4.6
	2–5	173	43.8
	6–10	107	27.1
	11–25	47	11.9
	26–50	29	7.3
	51–100	10	2.5
	> 100	10	2.5
Employm. Type[M]	Full-time employment	165	41.8
	Part-time employment	49	12.4
	Self-employed / freelancer	130	32.9
	Intern	30	7.6
	Student	15	3.8
	Hobbyist	124	31.4
	Unemployed	39	9.9
	Retired	3	0.8
	Other	6	1.5
Exp. w. Functionality[M]	Advertising	91	23.0
	Analytics	215	54.4
	Customer interaction	293	74.2
	Embedded media	258	65.3
	User login / authentication	318	80.5
	Payment	129	32.7
	Programming / design	328	83.0
	Privacy popups / forms	118	29.9
	Social media integration	204	51.6
	Website protection	130	32.9
Role(s) with Website[M]	Product / project manager	136	34.4
	Content creator / contrib.	142	35.9
	Social media manager	51	12.9
	Marketing	63	15.9
	Sales	19	4.8
	Quality assurance	93	23.5
	User experience	162	41.0
	(Web) developer etc.	337	85.3
	Admin / (web) operator	194	49.1
	Legal counsel	13	3.3
	Data protection officer	43	10.9
	Customer support/relations	71	18.0
	Other	19	4.8

5.4.1.2 *Websites Provided by Participants*

In Q2-0, we asked participants to provide a website they had recently worked on that would serve as a reference for Parts 3 and 4 of the survey. Data cleaning left us with 361 unique valid websites, for which we compiled descriptive statistics shown in Table 5.2.

The most frequently occurring TLDs were `.com`, `.org`, and `.de`, followed by domains associated with web development, such as `.github.io` or `.dev`. Thematic classifications by McAfee were available for 264 (83.8 %) domains, the most common being Business, Internet Services, and Education / Reference. 141 registered domains (44.8 %) appeared on the Tranco top 1-million list, with a mean ranking of 104,767 (min 5, max 958,899, SD 168,620.3, median 46,695). Overall we find that participants mainly reported international sites aimed at providing services or information, but also a significant amount of smaller and/or personal sites hosted on popular platforms and a multitude of other thematic categories, creating a diverse sample of websites.

Participants named 72 different countries as the seat of the company behind the website (Q2-3). Coding of the open-ended answers to Q2-4 revealed that the websites were mostly targeted at a global or multi-regional audience; Table 5.2 also lists the most popular individual target regions. Almost half of the websites (44.8 %) were reported not to have a website-specific revenue model (Q2-5). On average they relied on 0.91 sources of revenue (SD 1.03, min 0, max 5, median 1). Most common were products / services sold on websites (20.5 %), subscriptions / membership (17.5 %), and revenue streams not explicitly listed in Q2-5 (14.4 %).

Table 5.2: Statistics about the self-selected websites participants considered while answering the survey. [C] indicates coded open-ended answers, [M] indicates multiple-choice questions or multiply assigned codes or tags for which (response) counts can sum up to more than 100 %. Statistics in the left column are from Part 2 of the survey and percentage values are relative to the total number of survey responses (n = 395). Statistics in the right column result from the analysis of the website URLs provided by participants in Q2-0 and percentage values are relative to the number of unique entered domains (n = 361).

	Survey Responses	n	%
Team Size	I am the only team member	145	36.7
	2–5	141	35.7
	6–10	50	12.7
	11–25	36	9.1
	26–50	5	1.3
	51–100	5	1.3
	> 100	10	2.5
	Don't know	3	0.8
Country of Website HQ	United States of America	70	17.7
	Germany	46	11.6
	United Kingdom	21	5.3
	Russia	20	5.1
	Brazil	18	4.6
	India	15	3.8
	China	13	3.3
	Switzerland	12	3.0
	Canada	11	2.8
	The Netherlands	11	2.8
	Other	154	39.0
	N/A	4	1.0
Target Region/Audience[C]	Global	128	32.4
	Europe	56	14.2
	Multiple regions	30	7.6
	United States of America	26	6.6
	East Asia	17	4.3
	Brazil	15	3.8
	Southeast Asia	15	3.8
	Africa	12	3.0
	Russia / CIS	12	3.0
	North America	11	2.8
	Other	20	5.1
	N/A	53	13.4
Revenue model[M]	Targeted advertising	32	8.1
	Non-targeted advertising	22	5.6
	Affiliate marketing / links	21	5.3
	Donations	37	9.4
	Subscriptions / membership	69	17.5
	Sponsored posts / articles	22	5.6
	Products / services on website	81	20.5
	Other revenue streams	57	14.4
	None / not applicable	177	44.8
	Don't know	5	1.3
	Other	17	4.3
	N/A	2	0.5

	Website Analysis	n	%
Top-Level Domains	.com	107	29.6
	.org	30	8.3
	.de	24	6.6
	.github.io	19	5.3
	.herokuapp.com	17	4.7
	.dev	12	3.3
	.net	11	3.0
	.com.br	10	2.7
	.ru	10	2.7
	.io	7	1.9
	Other	115	31.9
Website Categories[M]	Business	65	18.0
	Internet Services	54	15.0
	Education / Reference	38	10.5
	Personal Pages	21	5.8
	Software / Hardware	19	5.3
	Interactive Web Apps	18	5.0
	Blogs / Wiki	15	4.2
	Marketing / Merch.	11	3.0
	Finance / Banking	10	2.8
	Online Shopping	10	2.8
	Other	129	35.7
	Uncategorized	48	13.3

Table 5.3: Reported functionalities on websites (Q2-6; n = 395), participants' involvement with them (Q2-7; relative to "present"), and, based on that, how often they were randomly assigned survey parts 3 and 4.

	present	Participants' involvement				assigned
		selection	integration	maintenance	none	
	n	%	%	%	%	n
Advertising	67	44.8	46.3	32.8	26.9	25
Analytics	251	47.4	40.6	46.2	17.1	126
Customer Interaction	268	53.0	46.6	45.1	10.8	138
Embedded Media	248	55.6	48.0	45.2	9.7	141
Login / Authentication	265	48.7	41.5	40.8	17.4	137
Payment	101	43.6	40.6	29.7	26.7	37
Programming / Design	355	61.7	57.7	46.2	8.7	235
Privacy Forms / Popups	136	40.4	36.0	33.8	30.9	57
Social Media	186	53.8	44.1	40.3	16.7	101
Website Protection	187	51.3	39.0	39.0	24.6	70

5.4.2 *Privacy Considerations in Selection*

To find out if privacy played a role in *the decision how to integrate* a desired functionality, we investigated what functionalities were present on participants' websites, whether they were integrated via first- or third-party solutions, and the underlying decision process, including considered alternatives, consulted information sources, and the people involved.

5.4.2.1 *Integrated Functionalities*

Survey responses

In Q2-6 we asked participants which of the ten functionalities identified in Section 5.3.1 were present on their website. Participants' websites used on average 5.2 of them (SD 2.3, min. 1, max. 10, median 5). In its "present" column, Table 5.3 lists how often each functionality was mentioned. The numbers show that the reported prevalence of the functionalities differs greatly. Most commonly used were programming or design resources (355 / 89.9 % of websites), customer interaction tools (268 / 67.8 %), and web analytics (251 / 63.5 %).

Website analysis

To assess the number of third parties the websites actually use, we combined the data collected from the three server locations (see Section 5.3.5) to ensure that no configurations dependent on visitors' IP address or region biased our results. Out of 361 unique websites provided we were not able to access 10. On average, each website contacted 6.2 third-party domains (min 0, max 144, SD 6.95, median 3) and 80 sites made no requests to third parties at all.

On 76 sites we observed third-party requests associated with functionality whose presence had not been reported in the corresponding response to Q2-6. The most common observation was a request to Google's advertising domain `doubleclick.com` (42 cases), followed by site analytics (14), CDNs (12), cus-

Table 5.4: Measured prevalence of common third-party services used on 351 websites compared to privacy-friendly alternatives.

Integration Solution	n	%
Analytics		
Google Analytics	158	45.0
Google Analytics w/ IP anonymization	24	6.8
Privacy-friendly (Matomo/Piwik)	15	4.3
Only privacy-friendly	11	3.1
Video		
YouTube	74	21.1
Vimeo	12	3.4
Privacy-friendly (YouTube-nocookie)	16	4.5
Only privacy-friendly	6	1.7
Maps		
Google Maps	38	10.8
Privacy-friendly (OpenStreetMap)	3	0.9
Only privacy-friendly	2	0.6
Design		
Google Fonts / Font Awesome	244	69.5
Privacy-friendly (3P-hosted)	6	1.7
Privacy-friendly (self-hosted)	86	24.5
Only self-hosted fonts	22	6.3
Programming		
jQuery from CDN	72	20.5
Privacy-friendly (self-hosted)	138	39.3
Only privacy-friendly	101	28.8

tomer interaction (6), and embedded media (5). The rest belonged to other functionalities not covered by the survey. The high prevalence of requests to advertising domains despite the fact that developers had not reported the use of advertising – confirmed by manual inspection – can be explained by third parties loading additional services [280]. The majority of these requests went to `doubleclick.com`, contacted by locally hosted Google Analytics scripts. Other cases involved social media bookmarking services like AddThis or ShareThis that contact various advertising domains.

In the other direction, 136 responses reported functionalities for which website analysis did not find obvious requests to matching third parties. The majority of these cases concern scripts for customer interaction (64), embedded media (70), or social media integration (46). Besides methodological limitations outlined in Section 5.5.5, the explanation was often that the functionality was hosted locally, for example, via CMS plugins, as reported in Section 5.4.2.2.

Prevalence of privacy-friendly alternatives

Last, we compared websites' hosting strategies against privacy-friendly recommendations [73]. Table 5.4 lists results for selected services. We found that for many common third-party services like analytics, videos, and maps the main strategy was to embed the well-known services. For example, 158 websites made use of Google Analytics, while only 15 used the privacy-friendly alternative Matomo. Out of those 15 another 4 were found to be using both, for example, on subsites. For more technical functionality like programming and design resources we observed more variation in first- vs. third-party hosting. While we found only six websites that used privacy-friendly font hosting sites (such as Fork Awesome or Fontello [73]), 86 hosted additional fonts on their own server. For the widely used web programming library jQuery the results were reversed: The majority (138) self-hosted the script, while 72 used CDNs to serve the files. Again there were sites that used both strategies, for example, when a library was used multiple times by different components or plugins.

5.4.2.2 *Prevalence of First-Party vs. Third-Party Solutions*

Q3-2 investigated how the different functionalities were integrated into websites. We focused on the hosting location (first-party solution, third-party software installed locally on the first-party system, or third-party service remotely included from vendor's server). For embedded media and social media, we also investigated (Q3-2c/2d) how remote resources were embedded into the website: via self-written code, code provided by the third party, or an embedding method provided by another third party (such as social media plugins that support multiple social media sites). Figure 5.2 shows the prevalence of each hosting and embedding type. For hosting (Figure 5.2 (a)) we observe that websites predominantly self-host solutions for customer interaction (user comments, contact forms, chat, etc.), privacy popups and forms, and embedded audio. Remotely hosted third-party solutions are dominant for analytics, payment, and hosting of embedded video and map content, while prevalence of the different hosting types was more varied in the other categories.

Local vs. remote hosting

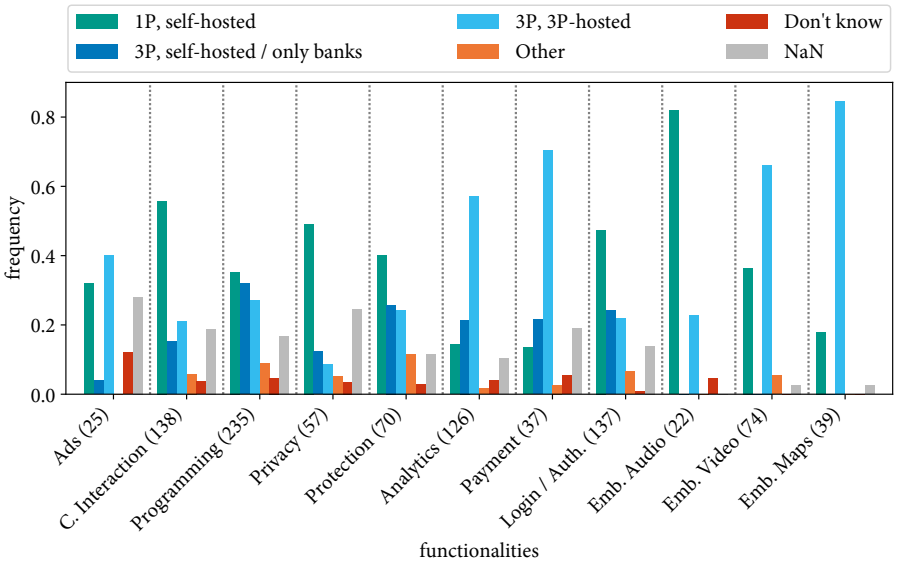
Embedding mechanisms

As shown in Figure 5.2 (b), remotely hosted media are typically embedded using the embedding code provided by the hosting service. Social media share buttons and embedded feeds, whose functionality implies the requirement to access an API provided by the social network, more or equally often use one of the two third-party embedding variants. By contrast, buttons or links to the website's social media profiles, which do not trigger an action specific to the social network, are more frequently integrated via first-party solutions.

Used third-party services

Q3-2 also asked participants to specify which concrete service the website used. Coding revealed the following categories of functionalities to have a clear market leader: advertising (Google Ads / AdSense / DoubleClick for Publishers [63.6 % of participants who used a third party and provided an answer]), analytics (Google Analytics, 65.7 %, followed by Matomo, 10.3 %), embedded videos (YouTube, 90 %), embedded maps (Google Maps, 62.5 %). We observed a more varied use of third-party services for programming and design resources (top 3: Bootstrap (18.2 %), React (17.5 %), jQuery (14.7 %)). For website protection, participants equally often mentioned web security libraries, which they considered self-hosted third-party services, and Google's reCAPTCHA as the most popular remote third-party service (12.1 % for both).

(a) Integration type (hosting)



(b) Integration type (embedding)

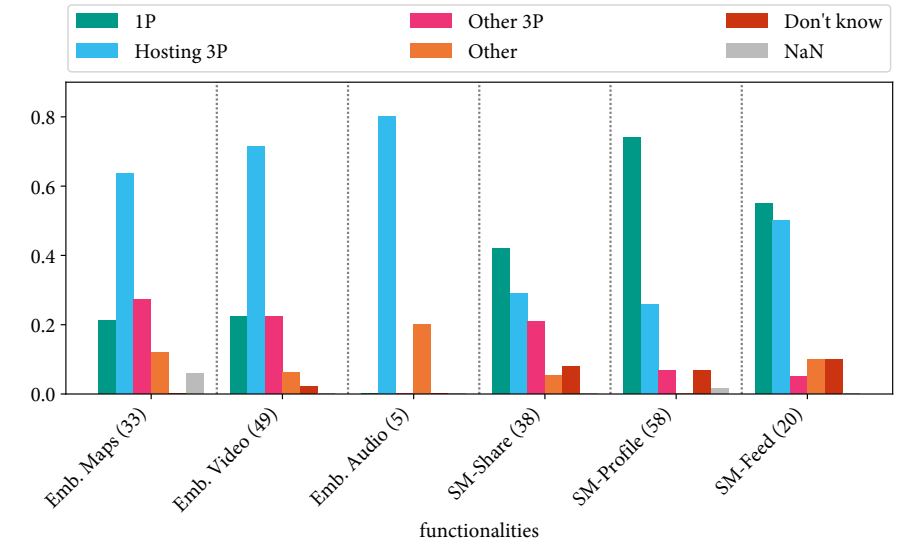


Figure 5.2: Integration type (Q3-2) for different types of website functionality. (a): use of first- vs. third-party hosting; (b): source of embedding code for embedded media and social media integration. Numbers are relative to how often the respective question had been displayed (see the survey logic in Appendix A.2). All n values are shown in the x-axis labels.

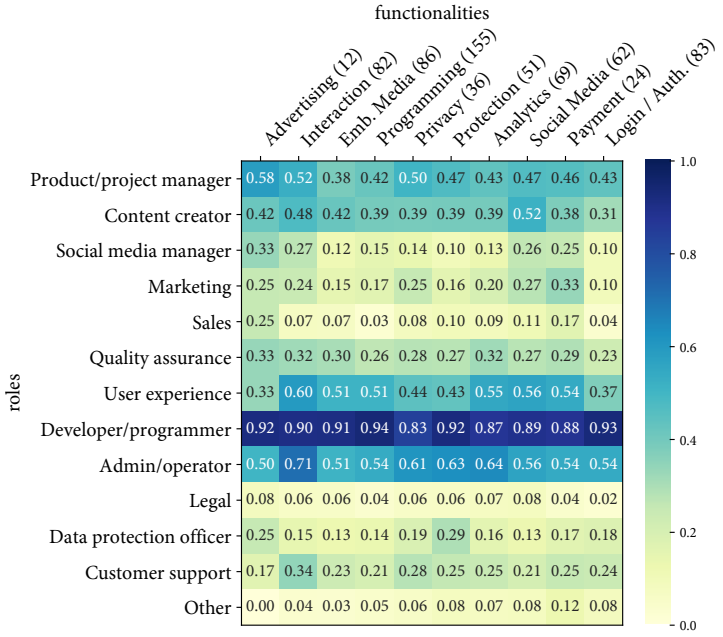


Figure 5.3: Responsibility for selection: for participants involved in the selection of a given functionality (Q2-7), their roles in relation to the website (Q2-1).

Overall, our findings match expectations: Third-party usage appears to be more prevalent for website functionalities that (mostly) require third parties to be involved, such as payment services or social media integration, or that previous work has assessed to be complex to self-host or implement, such as analytics or video and map resources [162]. As for the concrete third-party services used, web tracking research has repeatedly identified Google's services to be the most prevalent third-party services on the Web [63, 139, 282]. Still, we measured some efforts at privacy-friendly configuration of Google services.

5.4.2.3 *Decision Process*

Next, we investigated how people had arrived at these solutions to integrate different website functionalities.

PEOPLE INVOLVED IN THE SELECTION PROCESS We learned about who was involved in the selection process in two ways. For participants involved in the selection of how to integrate a functionality (Q2-7), we evaluated their roles with regard to the website (Q2-1). Figure 5.3 shows that across all categories, people involved in selection predominantly had technical roles. For given roles we also observed higher involvement in the selection of functionalities that closely relate to that role, such as customer support for customer interaction or sales for advertising. Q3-8 asked participants not involved in selection who had made that decision, with results shown in Figure 5.4. Here participants most frequently referred to developers, with the notable exception of privacy popups or forms, for which the decision often lay with the legal team, data protection officers, or management. This is also the functionality where participants reported the lowest involvement rates (see Table 5.3).

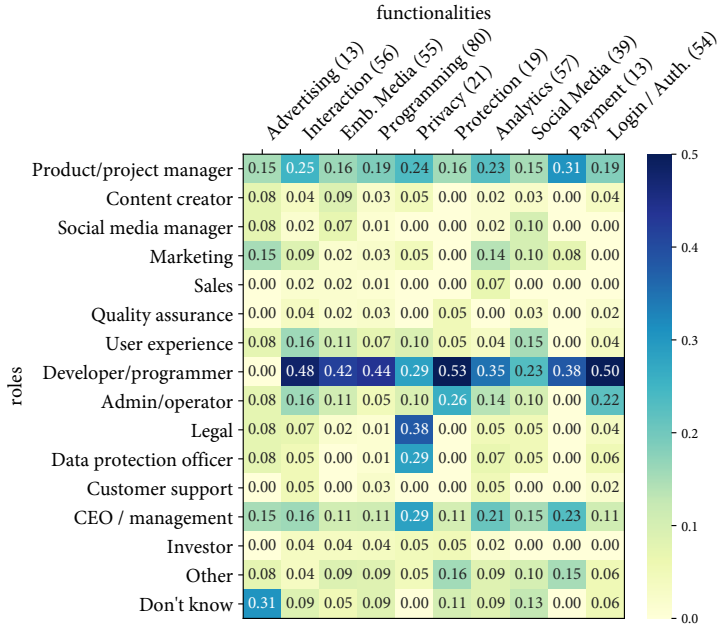


Figure 5.4: Responsibility for selection: for participants not involved in selection, who was responsible (Q3-8).

RESOURCES USED FOR SELECTION Across all categories, participants mainly relied on official websites and documentation to select how to integrate a given functionality (Q3-6); also frequently named were the website's team, online articles, and forums. The same information sources were reported as most commonly consulted in the selection of ad networks for mobile apps [183]. Also confirming the findings of previous work [15, 183], terms of service or privacy policies were rarely consulted, except for payment, privacy plugins, and advertising (16.7 % for each). Figure 5.5 has detailed numbers. This suggests that not even functionality where people directly enter sensitive information, such as customer interaction, prompts developers to look up a third-party service's data processing practices. This could be due to the complexity and length of these documents, which reinforces the need that third-party services present their key privacy practices in a condensed, easy to understand, and accessible form [15].

REASONS FOR THE SELECTION OF EXISTING SOLUTIONS Coding of the open-ended answers to Q3-3 identified reasons why the respective integration solutions had been selected for each functionality. Figure 5.6 investigates the reported reasons for two mutually exclusive groups: purely self-hosted solutions, whether first-party or via a locally hosted third party, where collected data is expected to stay on the website's host system, vs. solutions that only rely on remote third-party hosting and thus can involve information being sent to a third-party server. Figure 5.6 (a) shows the prevalence of each code for each of these integration types, aggregated across all functionalities. We find that the most prevalent decision factors for either integration type are ease of integration and features, though these play a bigger role in the adoption of pure third-party solutions. The "Other" category mainly comprises generic answers such as "I

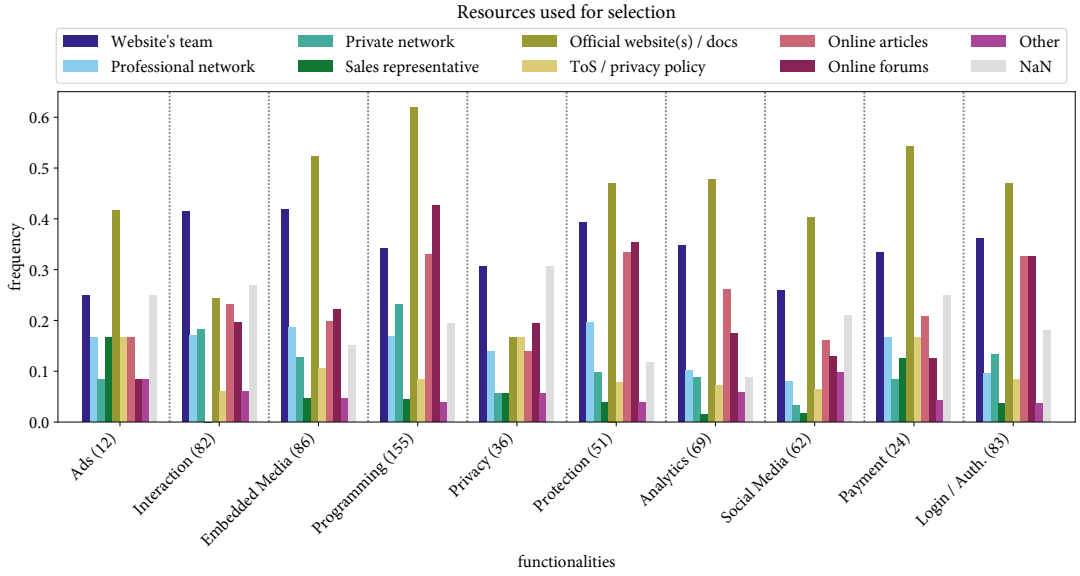


Figure 5.5: Resources used to select how to integrate a website functionality (Q3-6). Numbers are relative to the people involved in selection of the respective functionality, shown in the x-axis labels.

just like it" (P323-Social) or "it's the best" (P188-Login), which explains its relatively high prevalence.

Beyond these general factors for adoption, we observed that some mainly occurred for certain functionalities, such as revenue for advertising, legal considerations for privacy popups or forms, security for login / authentication, familiarity for programming / design and analytics, and popularity for payment. Privacy considerations were rarely mentioned, except for analytics ("I wanted something very minimalistic, non-intrusive" [P353-Analytics], "I care about users [sic] privacy" [P83-Analytics]). These observations confirm findings in the mobile space that third-party adoption is driven by the goal to save time and effort through code reuse [233] and additionally finds that these factors can fuel the reasoning both for or against third-party use and there are differences between functionalities.

Alternatives for hosting

CONSIDERATION OF ALTERNATIVES Participants involved in the selection of a functionality were asked in Q3-4 whether they had considered alternatives to their chosen integration solution. Figure 5.7 shows that across all categories, this was answered negatively by a large share of participants, from 16.7 % (advertising) to 50.7 % (analytics). A similarly low rate was reported in the work of Mhaidli et al. [183], who found only two out of nine interview participants to have made some effort in considering and comparing different mobile ad networks before settling on one. Rather, participants were found to select a network based on some "vague awareness" of what was popular and commonly used with good experience. We found similar sentiments in our data for functionality with a clear market leader, notably the prevalent use of analytics, for which the outstanding popularity of Google Analytics was confirmed by our measurements (see Table 5.4). The answers to Q3-2 suggest that people consider it the "default" solution and do not even think

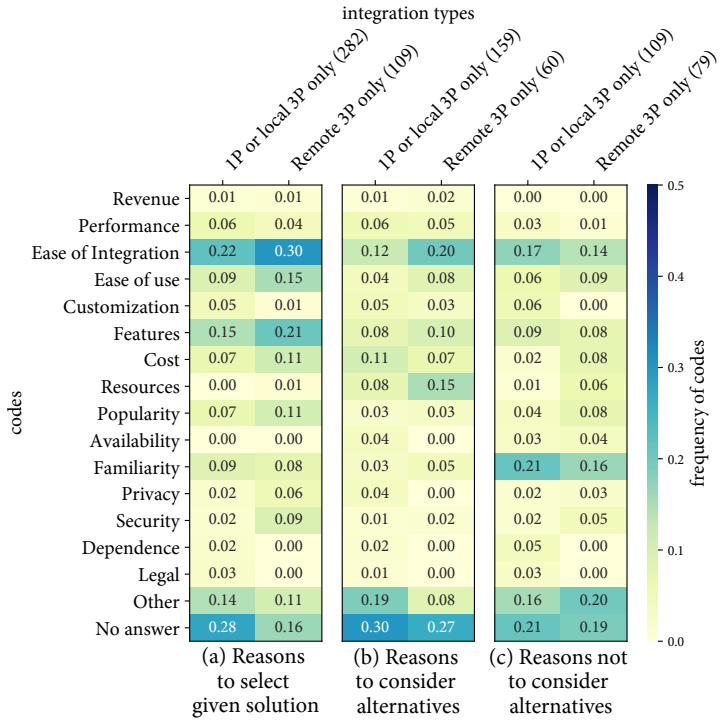


Figure 5.6: Reasons why a given website functionality was integrated in a certain way (a) and why alternatives were considered (b) or not (c), aggregated across functionalities.

about possible alternatives. Except for payment, which is only practical with the involvement of third parties, most considered alternatives were first-party solutions, even for functionalities considered difficult to self-host such as video content or (targeted) advertising [162]. This could again hint at people rarely choosing between different third-party services but rather deciding between either self-implementing a functionality or using a specific third-party service.

Alternatives for embedding

For embedded and social media, participants also had the option to indicate whether they had considered embedding mechanisms from other sources. Of the 62 people who had been asked this question for social media integration, 12 (19.4 %) had considered using code provided by the social networks and 4 (6.5 %) had considered code by another third party. The embedded media category was shown to 86 participants, 9 of whom (10.5 %) had considered self-written embedding code, 3 (3.5 %) code provided by the resource-hosting third party, and 4 (4.7 %) code by another third party.

Reasons (not) to consider alternatives

As for the reasons why alternatives were considered or not (Q3-5), Figure 5.6 in (b) and (c) investigates this for self-hosting vs. pure remote third-party use. We observe that, like for the selection of the current solution (a), ease of integration is a prominent factor to both consider and not to consider alternatives. Somewhat unexpectedly, for pure use of third parties this reason and resources appear to be factors to research rather than to not consider possible alternatives. This could hint towards users of third-party services not always being content with what those offer and decision processes to be complex. However, the most important factor not to consider alternatives appears to be familiarity with the selected solution, for self-hosted solutions even more so than for use of

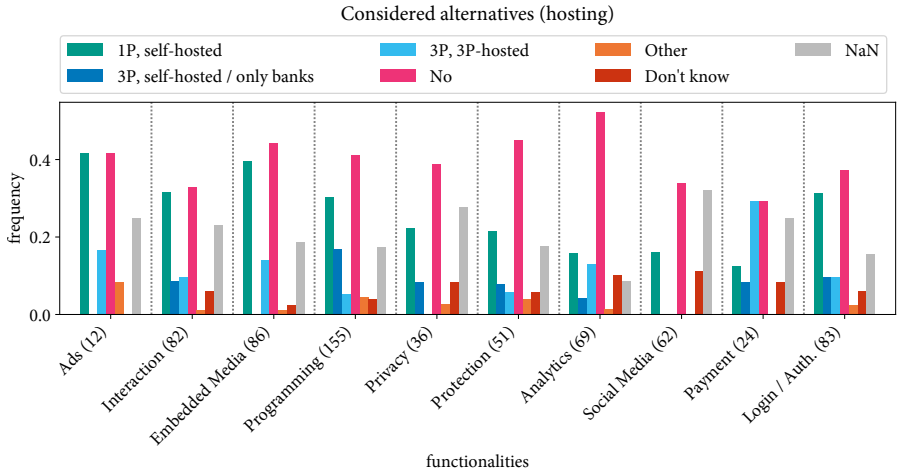


Figure 5.7: Alternatives considered (Q3-4) for the hosting of different types of website functionality. Numbers are relative to how often the question was displayed (see survey logic in Appendix A.2). All n values are shown in the x-axis labels.

remote third-party services. The "Other" responses to this question mainly comprised satisfaction with the current solution, low priority of the respective functionality, or mere statements that it was unnecessary to look for alternatives ("It wasn't required" [P316-Privacy]; "The first way worked" [P241-Analytics]).

5.4.3 *Privacy Considerations in Integration*

Beyond the selection phase, we investigated participants' privacy practices in the stage when the selected solution was actually integrated into the website. We asked participants for the resources they had used for integration and if they had made any specific efforts to protect visitors' privacy in the configuration of the integrated solution.

RESOURCES USED FOR INTEGRATION Figure 5.8 shows that the answers to Q3-7 paint a similar picture as the resources for selection (Q3-6). Again, the main sources of information were official websites / documentation and the website's team. Online articles and forums were less frequently used for actual integration compared to the selection phase. Terms of service and privacy policies again were rarely consulted. Though not directly comparable in answer space, the 20 % of privacy plugin users who consulted terms of service or a privacy policy are in the same dimension as the legal information sources used to integrate consent forms for advertising in mobile apps [265] (14.1 % for "Legal policies (e. g., GDPR)" and 9.9 % for "legal teams").

PRIVACY PROTECTION EFFORTS Q4-2 asked participants if they had employed specific measures to protect website visitors' privacy when configuring their solution to implement a functionality. Answers did not vary significantly across functionalities ($p > 0.05$, Fisher's exact test). For all of them, about a quarter of participants reported to have employed privacy protection mecha-

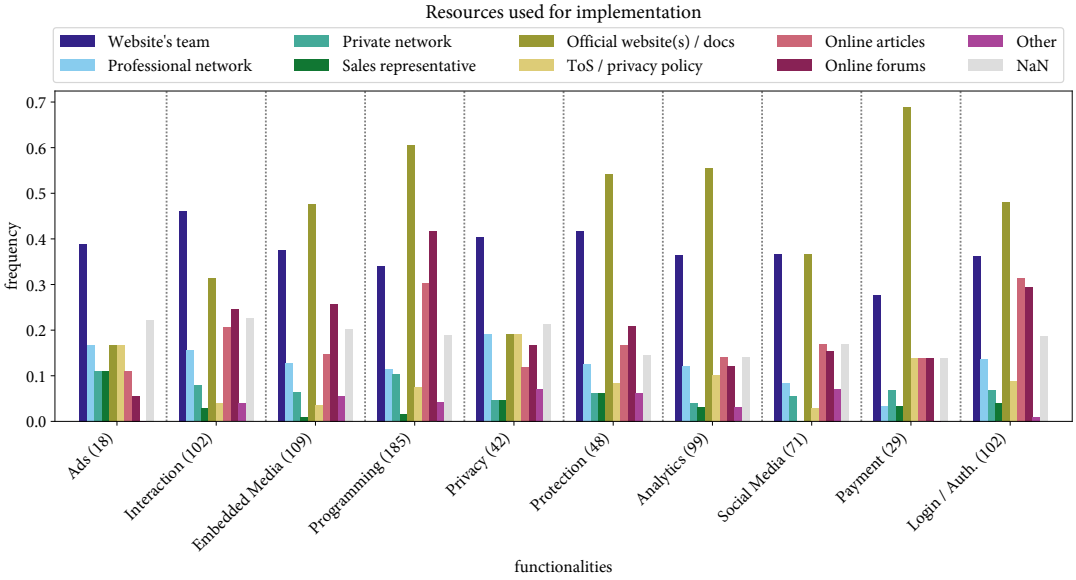


Figure 5.8: Resources used in the integration of a website functionality (Q3-7). Numbers are relative to the number of people involved in integration or maintenance of the respective functionality, shown in the x-axis labels.

nisms, another quarter stated to not have used them, about one third did not know, and the rest did not provide an answer.

Types of privacy protection efforts

Table 5.5 shows what privacy protection efforts participants reported to have made in the configuration of their solution. Participants frequently referred to data minimization ("I don't really collect user information, and when I do, I keep it to a minimum to get the job done" [P361-Programming]) and secure transfer ("encryption and [TLS]" [P84-Interaction]). Another prominent theme in the answers was first- vs. third-party selection, including self-hosting as a means to protect visitors' privacy ("Remove tracking from social media buttons by replacing them with a similar button" [P385-Social]), careful selection of the third party with privacy in mind ("I chose a font service that I believed would respect user privacy" [P136-Programming]), and using settings offered by the third-party service to collect less data. Prominent themes in individual categories are security for login / authentication (32.5 %) and customer interaction (28.1 %); anonymization, data minimization (22.2 % for both), and third-party settings (30.6 %) for analytics. The explanation for the repeated occurrence of security mechanisms, including access control, is that developers often conflate privacy with security [108, 269].

Reasons to protect visitor privacy

Across all categories, only 24 answers to Q4-3a also explained the motivation behind the measures to protect visitors' privacy. 20 named regulatory requirements mostly from privacy law but, in the case of payment providers, also due to industry regulations. Two participants mentioned an unspecified "requirement" for analytics and another two a self-commitment to privacy (for analytics and social media).

Reasons not to protect visitor privacy

Table 5.6 shows the reported reasons not to make privacy-protecting configurations. Most frequently, the solution was perceived not to collect any personal data, which was especially prevalent for programming / design (39.1 %; "be-

Table 5.5: Privacy protection efforts (Q4-2) reported by participants involved in integration or maintenance of a functionality, across all integrations (n = 224).

Code	Definition	Examples	n	%
No personal data	No personal data is collected.	"No user data is logged" (P337-Analytics), "No personal data is stored" (P46-Interaction)	9	4.0
Data minimization	Only the necessary personal data is collected; data collection is as minimal as possible.	"limited data retention" (P130-Protection), "only what we need" (P1178-Analytics)	38	17.0
Self-hosting	Services are self-hosted; all data stays within the respective organization.	"No external service used" (P190-Privacy), "Coded [it] myself safely" (P221-Social)	19	8.5
3P selection	Third-party services are carefully selected; there was a conscious decision for/against certain third parties.	"Remove GA :)" (P30-Analytics), "non-Google CDNs" (P855-Programming)	17	7.6
3P setting	Third-party services are configured in ways that increase privacy, e. g., by limiting the amount of collected data, encrypting data etc.	"use the no-cookie option" (P212-Embedded), "anonymize IP on [GA]" (P1256-Analytics)	26	11.6
User consent	Users were informed that their data would be available to third parties and gave their consented to this data processing before the functionality was loaded.	"I put them in containers [...] only executed after consent" (P214-Ads)	14	6.3
Transparency	Privacy policies or similar information on data practices are available to users.	"privacy policy" (P535-Ads), "we follow our privacy policy" (P66-Interaction)	4	1.8
Data access	The access to the data/server is limited; access is controlled.	"access to specific users" (P955-Programming), "don't pass any user data" (P191-Embedded)	18	8.0
Anonymization	Data is anonymized and cannot be used to identify certain individuals.	"anonymus [sic] identifiers" (P288-Analytics), "obfuscate user ids" (P917-Interaction)	12	5.4
Security	Security practices to avoid known attacks or vulnerabilities (e. g., to avoid XSS) are in place, that increase privacy by decreasing the probability of data leaks.	"HTTPS" (P855-Login), "password hash" (P619-Login), "encryption" (P1091-Interaction)	34	15.2
Other	Other concrete reasons not covered by the codes above.	"too many to list" (P163-Login), "look through the [...] source code" (P695-Embedded)	46	20.5
No answer	The participant did not provide an answer to the question, either by filling in nothing, something incomprehensible, or not providing an answer to the question.	Nothing entered, "1.?????????" (P352-Login), "Don't know specifics" (P53-Social)	29	12.9

Table 5.6: Reasons not to make any specific effort to protect visitors' privacy when integrating or maintaining a functionality, across all integrations (n = 263).

Code	Definition	Examples	n	%
No data collected	The solution does not collect any personal data, so there is no need for privacy protection.	"no tracking involved" (P213-Protection), "nothing is saved" (P247-Programming)	58	22.1
Data minimization	Only strictly necessary data is collected, so there was / is no need for privacy protection.	"We don't ask for anything beyond email address and name" (P44-Interaction)	8	3.0
Self-hosting	The service is self-hosted, and there is no need for additional measures as access is limited and no external services are involved.	"system is on-premise" (P201-Interaction), "own code without tracking" (P264-Social)	15	5.7
Trust in 3P	Trust in the third party to employ adequate measures to protect visitors' privacy.	"The service I used [...] handles security" (P380-Payment)	30	11.4
Impossible	Data collection cannot be controlled or limited, it is impossible to increase privacy.	"no configuration options" (P11-Ads), "we are not developing it" (P32-Protection)	23	8.7
Website purpose	The website's purpose makes privacy protection unnecessary, e. g., because its main content is only accessible in a logged-in state.	"Internal use only" (P95-Login), "page is not ready yet" (P361-Programming)	26	9.9
Priorities	Functionality (by adding third party services) has a higher priority than increasing privacy by avoiding these services.	"We [use] analytics to track users. That's the opposite of privacy" (P290-Analytics)	6	2.3
Payoff	Privacy measures include too much effort in terms of, e. g., workload, cost, time.	"It's more work" (P132-Analytics), "won't pay back" (P324-Privacy)	5	1.9
Unnecessary	It is not necessary to increase privacy. Answers with this code include no explanation, but often indicate a lack of awareness, care, or external requirements.	"Why should I" (P439-Social), "no need" (P63-Interaction), "Didn't have to" (P353-Programming)	38	14.4
Lack of knowledge	Participants are not able to adjust settings due to, e. g., a lack of knowledge or skill with the service.	"I can't understand whole of what [GA] collect[s]" (P382-Analytics)	2	0.8
Other	Other concrete reasons not covered by the codes above.	"Its just a frontend library" (P338-Programming), "Existing solutions satisfies" (P282-Programming)	16	6.1
No answer	The participant did not provide an answer to the question, either by filling in nothing, something incomprehensible, or not providing an answer to the question.	Nothing entered, "prefer not to say" (P91-Ads)	48	18.3

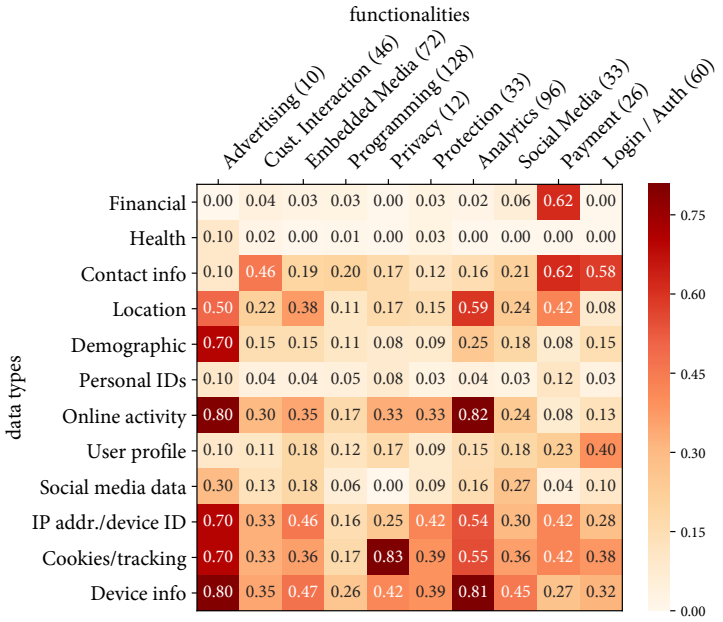


Figure 5.9: Percentage of third-party-using participants (n values next to functionality labels) who thought what types of personal data the service collected (Q4-1). Values are relative to the number of third-party users for the respective functionality, provided next to the x-axis labels.

cause the third party does not collect anything" [P109-Programming]), embedded media (18.6 %), and social media (34.8 %); in the latter case, the responses often referred to first-party integrations of profile buttons or links ("they're just links" [P98-Social], "simply images, wrapped in anchor tags" [P289-Social]). Other prominent themes were trust in the third party to adequately protect users' privacy ("I thought the default setup already protects the visitors' privacy enough" [P243-Analytics], "I trusted [Cloudflare] to not collect excessive information" [P321-Protection]) and the perception that it was impossible to do anything about collected data ("there is nothing I can do in GA to change the data Google collects" [P396-Analytics]), particularly for analytics (27.6 %). Trust in third-party vendors and the perceived inability to do something about the data collection were also recurring sentiments in why developers of mobile apps stick to a service's default configuration [183]. Finally, some answers simply deemed privacy protection unnecessary ("I don't care about privacy because 'data is king'" [P295-Payment]), prominently for programming / design (39.1 %) and embedded media (18.6 %).

5.4.4 *Awareness of Third-Party Data Collection*

Q4-1 more closely investigated the assumed lack of awareness of third-party data collection. For the third-party users of each functionality (as by Q3-2), Figure 5.9 shows the percentages who thought that the service collected specific types of data.

Types of data presumed to be collected

We observe that participants had a solid understanding of data collection implied by a service's core functionality. For example, a majority of participants reported that third-party privacy popups or forms collect cookies, that payment

services require contact and financial information, or that advertising and analytics collect device information and user online activities.

However, beyond this, participants' understanding of data collection was limited. This is especially evident in the case of IP addresses and device information: As HTTP(S) requests to a remote resource involve transmission of a user's IP address and user agent, this information is always available to the third party. More indirect is the opportunity for the third party to derive additional information via the collected technical parameters, such as tracking users across sites that use that service and learning about their browsing behavior. It appears that many participants embed third-party software and either do not know or are uncertain of the true extent of data collection by the third party.

Satisfaction with solution's privacy

This is supported by the responses to Q3-9 that let participants rate the integrated solution with regard to different metrics. Between 48 % (advertising) and 75.71 % (website protection) of participants reported to be *Satisfied* or *Very Satisfied* with the privacy offered by their integrated solution, while only up to 8.73 % (analytics) expressed some degree of dissatisfaction. This suggests that data collection by third parties is often either accepted or unknown.

5.5 DISCUSSION AND LIMITATIONS

Our findings provide insights into how web developers and people in similar roles *select* how to integrate a desired functionality, *configure* the selected solution, and if they are *aware* of the privacy risks associated with third-party services. For selection, we find the prevalence of third-party use to vary by functionality. In configuration, specific efforts to protect website visitors' privacy mostly appear to be made if mandated by technical guidelines on privacy law. Based on these findings, we discuss the need to raise awareness of the privacy risks of third-party use on websites and to promote adoption of privacy-friendlier alternatives. On the methodological level, our work is a case study for how the perception of research methods previously deemed acceptable can change over time. We conclude this section by discussing limitations of our findings.

5.5.1 *Lack of Awareness of Third-Party Data Collection*

Our research confirms the previously suggested lack of awareness [73] to what extent the use of third-party functionality on websites can pose risks to visitors' privacy. While developers appear to be aware of data collection closely tied to the main purpose of a third-party service, they often seem to not know or ignore the possibility that their visitors' personal data could be collected for other purposes or simply trust the third-party service not to collect data or to employ adequate privacy protection.

Legal requirements and guidelines as drivers of privacy protection

For analytics, our results hint at a somewhat higher privacy awareness than for other functionalities. This could be due to data collection simply being the main objective of web analytics or due to prominent and recent guidelines on GDPR-compliant use of web analytics [145, 263]. Similarly, concrete legal requirements in Article 5(3) of the ePrivacy Directive have led to the adoption of privacy notices and forms, while developers appear to find it difficult to

implement the more generic "data protection by design" approach promoted by the the GDPR or the NIST Privacy Framework [191]. Public discussion and additional guidelines could help raise awareness for the privacy risks of other types of third-party services on websites and provide guidance how to operationalize "privacy by design" for website development and integration, ideally addressing a wide range of website-related roles.

Measures to raise awareness would also need to communicate risks beyond the immediate control of developers, as third-party services often connect and share data with each other without users' knowledge, which leads to vast networks of third parties [118, 282]. Similarly in need of communication are different understandings of the sensitivity of the data collected in other contexts, such as IP addresses.

Aid through development tools

Referring developers to a service's privacy policy is insufficient to communicate its privacy risks. While privacy policies can be expected to contain information about the data collected by a third-party service, our results confirm previous work [15, 183] in that they are rarely used when selecting or configuring services. This is unsurprising given that privacy policies are notoriously hard to understand, and the GDPR, a law pursuing greater transparency, has even led to an increase in the length of online privacy policies [52]. In the mobile ecosystem, Apple's and Google's app stores have recently introduced additional aids in the form of *privacy labels* for mobile apps [9, 88]. With web development not taking place inside such a closed ecosystem, there are no centralized platforms developers could turn to for advice and comparison of different services that integrate a given functionality. For those who use common CMSes, their plugin repositories could introduce similar labels, placing privacy information more prominently than in a legal document. Alternatively, IDEs [158] and CMS editors could help assess the number of third-party requests in website code or problematic configurations for popular services and display advice.

5.5.2 *Promoting Privacy Engineering*

Developers' decisions matter

Our work confirms earlier findings from the mobile ad ecosystem that developers often feel resigned and unable to effect change in a third-party ecosystem governed by the exchange of revenue or functionality for access to website visitor's personal information: Previous work found sentiments that users' personal data would be collected by platforms and vendors, irrespective of the developer's decisions [183], and both developers [183] and third-party vendors [267] deem the respective other party responsible for the protection of users' data. One option to break this cycle of blame and instigate change would be to encourage developers that they can indeed make a difference through privacy-conscious integration of functionality [183]; after all, it is developers and end users that made these vendors that prevalent and powerful through use and promotion of their services. While in the past it was often browser vendors and developers of privacy-enhancing extensions who fueled advancements in website visitors' privacy, such as the option to block third-party cookies, relegating privacy protection to the browser comes at the risk of breaking websites and could overwhelm users with configuration options and prompts. Thus,

promoting privacy-by-design with website creators would be a more holistic approach that can ensure that privacy is considered from the beginning of the web development process, desired website functionality works as expected, and the burden is not placed on website visitors. A website that practices data minimization and privacy by design could even render annoying consent notices unnecessary for the benefit of both websites and visitors.

Considering privacy throughout the development process

We found notable involvement of data protection officers or legal experts only for privacy popups or forms, i. e., functionality added for the administration of a website's existing data processing practices. This could be an indicator that privacy is still regarded as something to be "added later" instead of being considered throughout the development process. Moreover, web development is often done in small teams or by single persons without a privacy professional at hand. When the decision is in the hands of developers and made in early stages of the development process, our results show that ease of integration and familiarity with solutions are the driving factors for adoption. This does not necessarily mean that developers do not care about privacy, but it is simply not an important concern given deadlines and limited resources in small teams [166]. While at the beginning of development it is often unclear what user data the final (web) application will need [15], this does not preclude the involvement of privacy considerations from the beginning. Iterative privacy impact and risk assessment processes that continuously evaluate functional requirements against privacy implications could help ensure that the desired functionality is implemented using the least amount of personal data, thus complying with frameworks that follow a data-minimization or privacy-by-default approach.

5.5.3 *Promoting Privacy-Friendlier Alternatives*

While advice to self-host [216, 225] or use privacy-friendly alternatives to popular third-party services [73] has increased in recent years, we found that only few participants heeded such advice. Others reported not to know alternatives to the used solution or did not have the time or resources to look for them. This should be interpreted as a challenge to better promote privacy-friendly alternatives for both the developers of these services and the privacy and security research community at large.

We found ease of integration, features, and cost to be among the most frequently reported factors that cause developers to adopt a certain solution – requirements currently easiest to satisfy by a service made available free of charge that instead relies on monetization of visitor data. It remains a major challenge to reconcile the demand for usability, features, and lowest possible cost if monetization of visitor data is not an option.

Enforcing privacy-friendly defaults

On the configuration level, privacy-friendlier options do exist but are often hidden or obscured by dark patterns [266]. For example, YouTube's setting for "privacy-enhanced mode" is only revealed when one scrolls down in the "Embed" dialog while the standard embed code is directly visible, and the documentation [100] only describes the standard embedding method in its main body of text and hides "privacy-enhanced mode" under an accordion menu. Vendors could encourage use of the privacy-friendly configuration by making it more prominent or even the default, though there is no incentive

for this if the service's business model is based on monetization of personal data, as is often the case with third-party services offered free of monetary cost. Privacy laws and court rulings were identified as drivers of privacy-related settings in ad networks [267], analytics services [198, 217], and cookie consent notices [69]. Thus, public policy measures and regulatory guidance could go one step further and require vendors to make the privacy-friendly option the default.

5.5.4 *Methodological Implications*

Section 5.3.4 described how recruitment via email addresses in public GitHub commit metadata came under the scrutiny of our state's data protection authority. Here we would like to discuss what part of the process had raised concerns with the DPA, what this means for future recruitment in privacy and security research, and what researchers could do in advance to decrease the likelihood of facing similar problems.

5.5.4.1 *Recruiting Developers on GitHub*

Visibility of email addresses on GitHub

Email recipients who asked how we found their email address on GitHub often pointed out that they had set their email address to "private" on their GitHub user profiles. While this setting hides the address from the public profile, it does not affect the visibility of the email address in commits to public GitHub repositories. Any given commit into a public repository has a corresponding `*.patch` file, available at `https://github.com/<user>/<repository>/commit/<commit_hash>.patch`. The second line in this file shows the author of the commit, along with their email address. This is due to the core concept behind GitHub's public repositories, where all commits, including metadata, are public. The documentation [95] describes how users can configure Git(Hub) to use their GitHub-provided "noreply" email address, which will remove their real email address from the commit metadata but still associate their contributions with their GitHub account.

Email feedback showed that many GitHub users are not aware of these mechanics and settings. This was also the issue at the core of the DPA's assessment, which argued that GitHub users pushing commits into public repositories did not expect to be contacted via their commit email addresses for the purpose of scientific research, and this lack of awareness constituted a legitimate interest of the user that outweighed public interest in scientific research. In addition, users of GitHub's API are bound by GitHub's terms of service and privacy statement [93]. GitHub's privacy policy considers a user email address public information (unless made private as described above) but proceeds to limit its use "for the purpose for which [the] user authorized it" [94]. Following the DPA's argument, this likely does not include being contacted for the purpose of participation in scientific research. It remains for the community to decide what influence such company policies should have on the question of what is considered ethical in privacy and security research, and, looking further ahead, how to handle company policies on data use that contradict what is permissible under applicable law.

Recommendations

For future recruitment of study participants we recommend, as also suggested by the DPA, to only use contact information that has visibly been made public by the individuals themselves with the intention of allowing the general public to contact them. GitHub's email address mechanics and users' lack of knowledge about them had neither been mentioned nor addressed by previous work that used public GitHub repositories for recruitment. We hope that our experience can inform the ongoing debate about ethics in privacy and security research and the search for alternatives to reach diverse sets of developers in a reliable, ethical, and affordable way.

5.5.4.2 *The Need for A Priori Community-Based Ethics Review*

Shortcomings of IRB review

It has long been best practice in human subjects research to obtain prior review via an institutional review board (IRB) or a similar entity to ensure that participation in the study does not cause undue harm to humans. However, in practice, many institutions, especially outside the US, do not have such a review board, or review is not always mandatory, as was the case for our study. But even if prior IRB review had been available, it remains doubtful whether it could have prevented the complaint to the DPA. The main goal of IRB review is to ensure that a study complies with human subjects regulations, not to provide a comprehensive ethics and legal assessment. In fact, we took additional steps to get GDPR assessments from our institutions' DPOs before running the study. The challenge is that in privacy and security research, a deep ethics and legal review would often require specific technical domain knowledge (e. g., GitHub's handling of commit email addresses), associated risks, and their legal evaluation. These are aspects that are often not covered by IRB guidelines or board members' background due to their differing function. Legal assessment in particular can be subject to rapid evolution through new laws and court rulings, requiring involvement of legal experts who keep up with this constant change.

Timing of ethics review

Recently the privacy and security research community has identified this need for thorough ethical review and multiple venues have set up ethics committees that can be involved in the review process if a submission raises ethical concerns with reviewers. This work went through this very process, and we highly value the thorough ethics review we received, which concluded that we adequately addressed our study's ethical implications. While ethical review after submission is an important step in ensuring that published privacy and security research did not cause undue harm to the people whose behavior and systems were studied, it effectively comes too late, at a time any potential harm would have already been caused. Hence, the community needs to consider how to provide ethical guidance before potentially harmful research is carried out, for example, by means of a "standing ethics review board" of expert volunteers that can complement institutional review in the study design phase. Such a priori ethics review would (1) help prevent unethical privacy and security research before it occurs, (2) provide researchers with experience and confidence in how to address ethical implications, and (3) minimize the sometimes arbitrary and ad-hoc assessments of a study's ethical implications by reviewers. An existing example is the Tor Research Safety Board [273]; providing committees of domain experts that cover the whole privacy and security field would pose a

major challenge. Hence, such a priori review would not have to be mandatory for all submissions but could become a valued community resource.

5.5.5 *Limitations*

Participant sample

Due to the complexity of web development and the nature of online surveys, our study has some limitations. First, we aimed to recruit a diverse participant sample and we are confident that it provides a wide range of perspectives on third-party adoption but may not include every type of website or third-party user. Websites and third-party services are not easy to categorize, and therefore participants might have interpreted our categories differently (see Section 2.1.5). However, we provided examples and aimed for a sensible compromise between lengthy explanations and too much room for interpretation.

Self-reported data and lack of compensation

Second, a limitation of any survey is self-reported data. We cannot verify to what degree participants were actually involved with the provided website or if they consistently answered for the same site. Analyzing self-reported information is common in research involving developers [1, 183, 213, 245], and manual inspection of survey responses suggests that participants answered consistently. Our survey was voluntary and uncompensated, which might have introduced bias, especially since experts tend to be well-paid and hard to reach. However, previous research found a lack of compensation to yield higher motivation or engagement in developer studies [1, 2, 102, 187].

Changes on websites

Further limitations apply to our website analysis. As data collection took multiple weeks, it is possible that in some instances websites changed between participants' responses and website analysis. Additional discrepancies might have been introduced due to our categorization differing slightly from WhoTracks.me, third-party vendors using the same domain for multiple services, or participants not knowing or naming the functionalities on their website.

5.6 CONCLUSION

In this chapter, we reported findings from an online study with 395 people involved in website development or maintenance on how common website functionalities are implemented, in particular whether third-party services are used and if and how respective privacy implications have been considered.

While we observe that the selection process is influenced by a variety of factors, we find that often factors such as a third-party service's popularity and ease of integration fuel adoption decisions. By contrast, website visitors' privacy only plays a notable role in web analytics, a functional category which has been explicitly addressed by data protection authorities. If alternatives are considered, the decision tends to be between a first-party integration and a specific third-party service, rather than between different third-party vendors.

Except for privacy popups and forms, data protection officers and legal counsels are rarely involved in the decision processes that lead to the integration of third-party services into websites despite potential privacy implications. As a potential reason we identified a widespread lack of awareness of data collection through third parties, especially regarding the transmission of IP addresses and device information, which can allow the third party to track people's browsing

behavior. Regulators and the research community are encouraged to raise awareness of the privacy implications of third-party use and find ways to assist website creators with privacy-friendly integrations.

6

NOTIFYING WEBSITES ABOUT NONCOMPLIANCE WITH THE GDPR

6.1 INTRODUCTION

Slow implementation of legal requirements in practice

The GDPR coming into effect has demonstrated that the adoption of new regulations and guidelines in practice is often slow, as shown in Chapters 3 and 4 for the integration of consent on websites and in Chapter 5 for the privacy-friendly selection and configuration of third-party services to implement the GDPR's "data protection by default and by design" principle. Our findings are supported by related work conducted around the same time that identified widespread deficits in the correct implementation of key GDPR principles including consent [174, 195, 234], transparency [163], and data subject rights [281]. While the last few years (as of November 2022) have seen increasing numbers of cases that resulted in fines being imposed under the GDPR [40], a lack of monetary and human resources continues to pose a problem in large-scale enforcement of privacy laws [111, 262]. Paired with the fact that web privacy research has been identifying privacy issues on websites at scale for years, this raises the question if and how the scientific community could aid regulators in identifying and remediating website behavior that violates privacy law.

Notification campaigns to help raise awareness

One promising means for privacy and security research to help increase GDPR compliance on the Web are large-scale email notification campaigns. Informing the operators of affected websites could help raise awareness of practices that do not comply with privacy law and encourage operators to fix the issue before they are subject to GDPR-mandated fines. Such notifications have been repeatedly used by security research over the last decade to raise awareness and motivate fixes of diverse issues including Heartbleed [59], Cross-Site Scripting (XSS) [260, 261], DDoS amplifiers [146, 299] and potential information leaks [157, 167]. If this approach turns out to also be viable for notifications about privacy compliance violations, this could take the burden off data protection authorities in enforcing existing laws and help website owners to better protect user privacy.

While both web privacy researchers and privacy NGOs have conducted notification campaigns about privacy issues before, these efforts have focused on selected Consent Management Platforms (CMPs) [197] or restricted the scope of their notifications to a single vendor and locale [169], typically because of the manual verification involved. It is also unknown how notifications about privacy problems compare to those about security vulnerabilities in terms of remediation rates and timing.

Our contribution: notifications about complex privacy issues

In this chapter we explore the feasibility of large-scale, automated email notification campaigns for vendor-independent violations of privacy laws, namely the GDPR's transparency requirement (see Section 2.2.1.2), its mandate to use state-of-the-art data protection mechanisms (Section 2.2.1.5), and consent to

the use of not strictly necessary cookies under the ePrivacy Directive (Section 2.2.2). To identify how notifications about privacy issues compare to those about security vulnerabilities, we also notify websites about publicly accessible Git repositories that may leak sensitive information. We compare fix rates between issues, investigate the impact of mentioning potential fines (see Section 2.2.1.6), and conduct qualitative analyses of feedback from recipient emails and a survey to learn how privacy notifications are perceived by recipients and what could be done to help them address privacy issues in the future.

More concretely, in this chapter we make the following contributions:

- We conduct the first large-scale, automated email notification study with 115K websites that investigates the feasibility of this approach for complex, vendor-independent privacy issues. Our notifications have significant impact on remediation rates for lack of a privacy policy or consent notice. We cannot identify any significant impact of warnings about potential fines.
- We compare the effect of notifications about privacy issues with those about a security vulnerability. For privacy, fix rates are lower than for the security vulnerability, which is also addressed more quickly. Recipients also perceive emails about a privacy compliance issue more negatively, partially because of a lack of intrinsic motivation to fix the issue or (incorrect) assumptions of the inapplicability of the relevant laws.
- To tackle the persisting problem of how to best reach web operators, we investigate if email addresses extracted from websites are an efficient and scalable alternative to prior approaches, which comprise manually collected addresses and email generics. Our results confirm this, leading to rates of 87.8 % and 33.8 % of successful handovers to recipients' mail servers for extracted addresses and email generics, respectively.

6.2 RELATED WORK

In this chapter we conduct the first large-scale, automated email notification study about complex, vendor-independent privacy issues and compare the impact of notifications about these issues to those about a security vulnerability. For this, we draw insights from prior notification studies about security and privacy issues and build upon mechanisms for automated detection of privacy problems.

6.2.1 *Security Notifications*

Notifications about abuse and security vulnerabilities

In web security research, large-scale notification campaigns were first used as a tool to alert web server operators about abuse of their infrastructure for a variety of unintended purposes, including distribution of malicious downloads [33, 286]. This approach was subsequently also applied to raise awareness of security vulnerabilities and motivate fixes for issues including Heartbleed [59], DNS zone poisoning [34], XSS vulnerabilities [261], HTTPS misconfigurations [299], DDoS amplifiers [146, 157], misconfigured IPv6 firewalls [157], and leaks of information whose public accessibility could pose a security risk, including industrial

control systems [157], Git or Apache Subversion (SVN) repositories [167, 260], cryptographic keys, database backups, server status information, and `phpinfo` files [167].

The reachability problem

One of the core problems in conducting large-scale web security notifications is to reliably reach the people responsible for fixing the issue. While previous work has found that more individual communication channels such as telephone [260], physical mail [167, 169, 260], contact forms on websites or associated social media accounts [260], or contact email addresses manually identified on the problematic website [167, 169, 260] can lead to higher delivery rates, the involved overhead in terms of human resources and monetary cost makes these infeasible for notifying websites at scale. Hence, most web security notification campaigns have used generic approaches to contact websites via email, either directly trying to contact the website owner(s) or operator(s) via generic email addresses [33] such as `info@DOMAIN` or `webmaster@DOMAIN` (RFC 2142 [46]) or WHOIS contact information [33, 34, 59, 157, 260, 261, 299] or through trusted third parties including CERTs [157, 261], DNS nameserver operators [34], or hosting providers [33], depending on the investigated issue(s). Drawbacks of this approach include high bounce rates due to missing or outdated information in WHOIS records or non-use of RFC 2142 mailbox names [34, 261]. The use of intermediaries carries the risk of them not forwarding notifications [157].

Message authenticity and content

Even if a notification email is correctly targeted, there still is the problem of it being considered spam or otherwise malicious by both mail servers and human recipients. Prior work has studied how to increase perceived message authenticity in notification campaigns by evaluating the effect of sender reputation [33, 169, 299], email format such as plaintext or HTML [260], text localization [157, 299], and use of S/MIME [260], but no clear "recipe" has emerged. Finally, findings also differ for the content of the notification message itself: While some studies did not identify a significant influence of message text on remediation rates [34, 299], others found that more detailed explanations had a significant positive influence [157, 286]. The tone of the message was found not to affect fix rates [260, 299]. Maass et al. compared existing work and outlined practical recommendations for future notification studies [168].

6.2.2 *Privacy Notifications*

Requirements for privacy notification studies

Compared to abuse and security notifications, previous work that notified website owners about privacy issues at scale is scarce. Challenges lie in such a study requiring

1) determining if the examined website is subject to the privacy legislation of interest,
2) certainty that a given issue is regarded a violation of this privacy legislation, and
3) a high-accuracy detection mechanism to keep the number of false positives low and not cause unnecessary anxiety and costly investigations with people whose websites do not have a privacy problem.

The third prerequisite is challenging, as privacy issues are hard to detect automatically [236] unless focused on specific services or vendors. In the latter case, an underlying common implementation can allow for simple yes / no checks of a value, parameter, or URL. Otherwise, complex heuristics are necessary that may require, for example, contextual analysis and natural language processing (NLP) to determine if a privacy notice contains the required disclosures.

Vendor-specific privacy notifications

Consequently, prior privacy notification campaigns focused on specific vendors or consent frameworks. Maass et al. [169] notified the owners of 4,754 German websites about the lack of IP anonymization in their Google Analytics integration. This is a configuration which the German DPA had deemed necessary for GDPR compliance [145] and can be remotely detected with certainty through URL parameters passed in the HTTP request to Google. While the study found that framing notifications as legal compliance issues led to increased remediation rates, the analysis only comprised German sites and those with an imprint, thereby limiting its insights to this selected group. Our work uses a significantly larger domain set without this bias.

In May 2021, Austrian privacy NGO noyb ("none of your business") notified more than 500 companies about consent notices on their websites that used techniques considered to be non-GDPR-compliant by various national DPAs [197]: consent banners without a "reject" option on the first layer, "reject" options presented as a link instead of a button, deceptive button contrast or color, pre-ticked checkboxes, justifying data collection with legitimate interest (Article 6(1)(f) GDPR), categorizing non-essential cookies as essential, and making it harder to withdraw than to give consent. While 42 % of individual violations were fixed within 30 days, 82 % of companies still exhibited some type of GDPR compliance issue. Since the goal of the notification campaign was to file complaints to national DPAs, the automated detection mechanism was targeted at a single CMP, OneTrust [204], and supported by manual review by legal experts, limiting this approach in coverage and scalability.

Our approach

By contrast, our work uses more exhaustive checks to detect privacy issues independent from a given vendor or framework. For example, we combine metrics from official CMP lists and banner-blocking browser plugins to automatically detect consent notices at scale. This also allows us to determine if notifications about such issues lead to remediation that involve more than just changing a single line of code, as in the activation of IP anonymization in Google Analytics.

6.2.3 *Automated Detection of Web Privacy Issues*

Automatically detecting vendor-agnostic privacy problems on websites is challenging due to a lack of standardization and concrete guidelines by lawmakers, data protection authorities, and court rulings on how to implement key legal requirements on a technical level, including transparency mechanisms mandated by privacy law such as privacy policies or consent notices. This is partly a deliberate decision to remain flexible towards future technological developments [236]. But even for concrete requirements, such as the wording of the "Do Not Sell My Personal Information" link mandated by the CCPA (see Section 2.2.4), actual implementations on websites widely vary [199, 285].

Automatic detection of transparency mechanisms

Still, there is previous work that worked towards automatic detection of the privacy issues at the heart of our study. A growing body of literature has tackled the problem to automatically find privacy policies on websites and download them for further analysis. Hosseini et al. [117] discuss and evaluate different approaches and identify best practices.

Recently web privacy researchers have also shown growing interest in automatically detecting consent notices on websites. As with privacy policies, differences in implementation make automatic detection difficult [236]. Thus, past work has focused on specific consent frameworks or individual CMPs [19, 114, 174, 195, 197] or included manual analysis, as in our works presented in Chapters 3 and 4. Despite EU law requirements that consent to data processing must be prior, free, specific, informed, unambiguous, readable, accessible, and revocable [14, 236], large percentages of consent notices on websites were found not to offer sufficient choice (again, see our work in Chapters 3 and 4), use dark patterns to nudge people into giving consent [195], or do not have a backend which ensures that visitors' selection is honored by the website [19, 174]. Bollinger et al. [19] trained a machine learning classifier on cookie–purpose mappings from CMP classifications and manual categorization by web developers. Examining a set of 29,398 websites from the Tranco top one million websites ranking that featured one of the investigated CMPs with cookie–purpose mappings, they found that 94.7 % exhibited at least one violation of a consent requirement. Like prior work that aimed to determine the purpose for which specific cookies are set [119], their cookie categorization approach suffers from a lack of reliable ground truth.

In this chapter, we build upon some of these techniques to automatically detect privacy issues at scale and independent of software or vendor, focusing on keeping the number of false positives low to avoid unnecessary notification of compliant websites. In the following (Section 6.3), we describe our measurement and notification infrastructure, along with associated ethical considerations and limitations. Section 6.4 reports on the measurement results of our campaign. In Section 6.5, we then describe our analyses of email communication and survey responses and present recipient feedback and survey results in Section 6.6.

6.3 MEASUREMENT AND NOTIFICATION SETUP

Our study setup first required us to identify the security and privacy issues we wanted to notify operators about and how to check their presence at scale. Further, we describe how we created a set of domains to monitor, the notification messages, and report infrastructure. We also explain ethical aspects of our large-scale measurements and notifications and discuss limitations of this approach.

6.3.1 *Investigated Issues and Implemented Checks*

Issue selection and check infrastructure

In Section 6.2 we already identified three core requirements for privacy issues to investigate in a large-scale notification study. Of particular importance is a low number of false positive cases to not erroneously alert recipients and

potentially trigger costly and stressful investigations. Thus, for privacy notifications, we worked with a legal expert on data protection law to identify issues that constituted a clear violation of unambiguous regulatory data protection requirements and could be automatically detected. This ruled out issues requiring human judgment, such as dark patterns. To estimate the prevalence of false positives for our checks, we manually verified, for each issue, whether it was indeed present on 250 websites randomly drawn from the set of all domains we found to have the respective problem. We ended up selecting four privacy issues which fit our requirements and implemented them as custom functions in OpenWPM [63], the web privacy measurement framework already used earlier in this work. To compare the effect of privacy notifications to those about a security vulnerability, we also selected one security issue already used in prior notification studies, publicly accessible Git repositories [167, 260]. For performance reasons, checks for the Git issue were not conducted with OpenWPM but with standard HTTP requests.

All checks were performed once a day, launched shortly after midnight CET from CISPA servers on the premises of Saarland University in Saarbrücken, Germany. Performing the checks with an IP address in the EU is important because some websites, particularly with `.com` TLDs, show consent notices only to EU-based visitors [284].

6.3.1.1 *No privacy policy*

As established in Section 2.2.1.2, the GDPR requires for all processing of personal data that the data subject is informed about the processing. Personal data also comprises communications data such as users' IP addresses in web server logs [67], even if no additional information is collected or the logging is only temporary. Thus, any website collecting such information must have a privacy policy that explains the use of visitors' personal data.

Finding privacy policies on websites

To determine if a website had a privacy policy, we followed the best practices identified by Hosseini et al. [117] and searched for privacy-policy-specific words in and around HTML link tags. For this, we extended the list of common words for privacy policy links, terms of service, and contact pages in all official EU languages we had already used for website-based recruitment in our study presented in Chapter 5[1]. After a website had been fully loaded, we used the list of words identifying privacy policies to find links that likely lead to a privacy policy. If no such link was found, we also visited less privacy-specific subpages like terms-of-service and contact pages and searched them for words from the privacy list. If neither of these searches led to a match, the site was marked as violating the privacy policy requirement. For the manually checked sample of 250 sites drawn from those with this violation, 0.4 % were false positives.

1 See Section 5.3.3 for how the list was originally created; even though that study required phrases for a slightly different set of languages, we had also compiled phrases for all other official EU languages back then. For the notification study, we also moved terms referring to pages with general legal information to a distinct list, which we had not done in the earlier study.

6.3.1.2 *Use of third-party cookies without consent notice (*No Consent*) or before interaction with consent notice (*Before Consent*)*

Legal basis: ePrivacy Directive

In Section 2.2.2 we already described that the act of setting and accessing cookies in users' browsers is regulated by the ePrivacy Directive and its respective implementations into national laws, as this is the more specific legislation on matters of electronic communication (see Section 2.2.3). As a reminder, under its Article 5(3) the storing of information in a user's terminal equipment, including HTTP cookies that are not strictly necessary for the functioning of the website, is only allowed if the user has given prior, active consent. As also mentioned earlier, mere passive access or continued browsing of the website do not constitute valid consent under EU privacy law [69, 71].

Detecting consent notices and problematic cookies

For the *No Consent* and *Before Consent* issues we focused on cookies set by third-party providers for advertising, analytics, and social media, because European DPAs had universally deemed these non-essential for the website's functioning [11, 135]. We used the WhoTracks.me database [91] to categorize a checked website's third-party requests by purpose and flagged those that included a `Set-Cookie` HTTP header and requested a third-party domain classified as "audio video player," "ad/pornvertising," "site analytics," or "social media."

The presence of a cookie consent notice was determined based on two rule sets: a list of common HTML elements from the EasyList Cookie List [61] and the list of Consent Management Platforms (CMPs) vetted by IAB Europe's CMP Compliance Programme [127]. If one of the EasyList rules matched or a script referring to one of the CMPs was found on the front page, we assumed that the site had a cookie consent notice.

Two cases of faulty consent

If a website issued a third-party request that required prior consent but a consent notice was not detected, we considered the site a case of *No Consent*. If a consent notice was detected but the flagged request was issued despite our script not interacting with the website, the website was labeled as having the *Before Consent* issue. In our manual check the prevalence of false positive cases was was 2 % for *No Consent* and 6.8 % for *Before Consent*. The latter involved a tradeoff between false negatives for the presence of a consent notice and false positives for a notice without a working consent mechanism.

6.3.1.3 *Input fields for personal information without HTTPS (*No HTTPS*)*

In Section 2.2.1.5 we described how the GDPR's requirements for "data protection by design and by default" (Article 25) and "security of processing" (Article 32) mandate the use of appropriate state-of-the-art technology for the collection and processing of personal data, which include the use of HTTPS.

Identifying input fields for personal data

To detect if a website requested users' personal data without securing it with HTTPS, we created a list of terms that described personal information (e. g., `firstname` or `password`) and were likely to be used as names for input fields that request the corresponding piece of information. In an iterative process we checked our list against the actual names of form fields used by websites, removed terms that led to many false positives, and added newly found, more specific terms (e. g., `login_email`). We ended up with a final list of 24 phrases that is not comprehensive but designed to reduce false positives: `newslet-`

ter, login, email, username, e-mail, name, firstname, lastname, gender, birthdate, bday, dob, dateofbirth, sso, signin, signin__email, login_email, loginmodel-username, connection_mail, email_address, login-user, user_login, email_1_db, login_pwd_db.

We flagged a site as violating the HTTPS requirement if one of the terms on the list was used in the name or id attribute of HTML input fields and the site did not use HTTPS. Manual validation yielded a prevalence of false positive cases of 3.6 % on the 250 sampled sites with this issue.

6.3.1.4 *Publicly accessible Git repository (*Git*)*

Security issue selection

If repositories for software version control systems such as Git or SVN are accidentally publicly accessible, they could potentially leak confidential information to outsiders, such as hardcoded encryption keys or credentials [260]. Considered a security vulnerability, this issue was already the subject of previous security notification campaigns [167, 260]. We selected it as a security issue for comparison against our privacy notifications because it is still a common problem, can be accurately detected, and, like the privacy issues, concerns a specific *domain* rather than a specific server (that could host multiple domains). Most importantly, this issue can be tested in a non-intrusive manner, which is a core aspect of ethical security vulnerability checks [260, 261] and excludes any vulnerability for which even a proof-of-concept would require some server-side code execution, which could be considered illegal in some countries.

Implemented check

To check websites for publicly accessible Git repositories, we used standard HTTP requests, as they were faster and less resource intensive than OpenWPM. We tried to access the file domain.tld/.git/config; if it contained the line [core], we requested .git/HEAD. This either directly provided the hash of the currently checked out commit or pointed to a branch, so we could retrieve that branch's commit hash from .git/refs/heads/<branch>.

If the commit hash could also be found on GitHub, we did not classify the domain as problematic, assuming that public availability of a repository already published elsewhere does not increase the attack surface [260]. Our check did not further investigate if the repository indeed posed a security risk, because once the presence of a publicly accessible repository has been confirmed, it would be unethical to search its content for sensitive information.

6.3.2 *Initial Domain Set*

In order to obtain a large and diverse initial set of websites to analyze, we leveraged a public domain list provided by the TheInternetBackup project[2], [23], whose goal was to compile a list of every domain on the Internet. To this end, users could upload domain lists that were only checked for a valid DNS result. Our starting point was a domain list with 252 million domains from February 2020. To reduce the number of sites subject to resource-intensive checks, we defined additional criteria for a website to be a candidate for a detailed check:

2 The site no longer exists, but a previous version can be accessed via the Internet Archive: https://web.archive.org/web/20200110214540/https://theinternetbackup.com/#about.

Criteria for candidate domains

- *EU-based*: To ensure that EU privacy laws applied to the monitored websites, we first resolved the domain names and checked that all requested IP addresses were within the EU, based on Maxmind's GeoIP database [175].
- *Not parked*: Next, we excluded domains for which the resolving nameserver was a known domain parking service. We identified these by manually extending the list by Vissers et al. [287]. These DNS-based checks reduced the number of candidate sites to 51 million.
- *Active web server*: We issued HTTP requests to all remaining domains to check whether they provided a website. If the HTTP response status was below 400, we kept the domain in our data set, leading to around 30 million candidate sites.
- *No previous opt-out:* We further excluded 1,513 websites that had asked to be removed from prior notification studies [260, 261] conducted by one of the authors of the conference paper that served as the basis for this chapter.
- *Public audience*: We excluded sites that only offered limited content or did not seem to be targeted at a public audience (e. g., "under construction" sites). As a metric we required that a website had at least five same-site links on the front page, otherwise we removed it from the domain set.

Pre-study to assess prevalence of issues

The check for internal links was part of a pre-study in which we visited all 30 million sites with our OpenWPM-based check infrastructure. It took three months to visit each domain once with our automated setup and check it for the presence of the four privacy issues. Overall we found 6,272,813 candidate sites (~21 %) with at least one privacy issue. Cases were not evenly distributed; most common was *No Privacy Policy* (17.44 %), followed by *Before Consent* (7.57 %) and *No Consent* (7.34 %). *No HTTPS* was rarest, with 2.85 % of sites.

Domain sample

Checking all of these domains daily, let alone notifying all of them would have been infeasible with the given hardware restrictions, so we sampled 500,000 domains from the list of domains with at least one problem. Then, for each issue, we sampled up to 100,000 domains; since only about 1 in 10 problematic domains had the *No HTTPS* problem, we only drew around 45,000 domains for this issue. In total, this left us with 331,222 domains with privacy issues subject to further monitoring. While this sampling was done in late September 2021, the set we sampled from contained all domains that had been identified as problematic once within the three preceding months. In addition, we found 58,715 domains with the *Git* issue, which we also added to the set of domains to check each day. In total, this yielded 388,825 domains for further consideration.

6.3.3 *Notification Emails and Infrastructure*

The notification process itself involved determining the email addresses to contact, the mail server setup, composition of the notification emails, and setting up a website that allowed participants to learn the current check results of their website and general information about our study.

6.3.3.1 *Contacted Email Addresses*

Automatically finding email addresses on websites

To identify points of contact with the monitored websites, we investigated a potential alternative to manually identified or generic email addresses: automatically finding email addresses on websites. As part of our daily OpenWPM checks, we searched the assessed websites for email addresses likely to belong to people involved in the website's technical or legal administration. For this, we identified links to privacy policy pages as described in Section 6.3.1.1. Using a regular expression, we searched the HTML code of these subpages for email addresses. If none were found on a privacy policy page, we extended the search to subpages expected to contain generic contact information, such as "About" or "Contact us" pages. To remove false positives (e. g., file names containing the `@` character) and to prevent emailing someone unrelated to the website we wanted to contact, we used only addresses whose domain name matched that of the website to notify. If this procedure yielded at least one email address for the inspected domain, we emailed up to three discovered email addresses in the order served by our database and flagged the domain as being notified through (a) *Parsed* email address(es).

Email generics

If no email address meeting the above criteria could be found, the domain was flagged as *Generic* and we sent our notifications to three generic aliases (RFC 2142 [46]): `info@DOMAIN`, the most frequently found email address in the first step, plus `webmaster@DOMAIN` and `abuse@DOMAIN`, the two most commonly used email generics according to the findings of Soussi et al. [252].

6.3.3.2 *Mail Server Setup*

To send notification emails, we used a designated server (`notify.cispa.de`) outside CISPA, i. e., hosted with an external server provider. This reduced the risk that our notifications negatively impacted our institution's normal email communication (e. g., in case we hit a spam trap). Both A and MX record of our subdomain `notify.cispa.de` pointed to this server. This subdomain was also used in the EHLO message. The server configuration followed best practices to increase the delivery rate, including SPF and DMARC records and DKIM signatures for outgoing emails. The policy in the DMARC record was set to 'none,' the percentage to 100 %, and the address for aggregated reports to `administration@notify.cispa.de`. We also configured the reverse DNS to point to `notify.cispa.de` to create another clear connection to our institution and S/MIME-signed all emails to enable validation by the receivers' email software. Finally, to reduce strain on receiving servers, we set the rate of delivery to our MX to at most one email every two seconds.

6.3.3.3 *Notification Emails*

In our notification emails we openly identified ourselves as researchers and the purpose of the emails as being a scientific study. The sender name was composed of the name of the researcher who was responsible for the notification setup, Matthias Michels, and his institution, CISPA. Mails were sent from a designated email address, `notify@notify.cispa.de`. The emails' subject line was "[Security and] data protection issue[s] on your website [DOMAIN]"

or "Security issue on your website [DOMAIN]", depending on the type of detected issues. Following prior findings that language did not have a significant influence on fix rates [299] and localization of notification messages may even increase the likelihood that recipients perceive them as malicious [157], all emails were sent in English. The message body introduced our project, the involved research groups and institutions, and the security and / or privacy issues we had identified on the respective website. We provided a description of the problem(s) and why they constituted a violation of privacy law or a potential security vulnerability. Appendix A.3 contains an example notification email with all possible issues.

6.3.3.4 *Report Website*

To aid operators in fixing their websites, we followed prior work [167, 261] in providing a web interface that allowed them to track their website's status with regard to the investigated issues. Every email contained a link to a domain-specific online report, which again listed and described the issues we had found on the respective website but was updated daily with the most recent check results. This allowed operators to learn if our checks still detected the corresponding issue or considered it fixed. In order to prevent incorrect feedback to operators due to a flaky check, an issue's state was only reported as fixed if this was supported by the latest two checks.

The online report also provided operators with the option to exclude their website from our checks and, for each detected issue, a form to report false positives. A screenshot of an example report is shown in Figure 6.1.

The website serving the online reports was hosted at CISPA, from where the daily checks were conducted. It also contained an introduction to our research project (see Appendix A.4), an imprint, and a privacy policy explaining our data processing.

6.3.4 *Research Ethics*

IRB approval and data minimization

To ensure our research followed ethical best practices, we requested approval from CISPA's IRB. We outlined that our measurement setup would not collect personal data beyond what was publicly available, i. e., email addresses found on websites or generic aliases. Beyond that we followed data minimization principles: Survey responses (see Section 6.5) were anonymized and did not contain any information that allowed us to identify the website or email address used for notification. The only information passed to the survey via URL parameters were the issues found on the website, notification round (initial notification, reminder, or control group debriefing), email type (*Parsed* or *Generic*), and study condition (*Warning*, *No Warning*, or *Control*); for all but the first, see Section 6.4.1. We received IRB approval without any requested changes to the study protocol.

Best practices for network checks and notification studies

In addition, we followed best practices recommended by prior work for ethical network checks [58] and notification studies [168]. We communicated our identities and benign intentions at all points of contact with the monitored websites and their operators: In all notification emails and on the study website (see Appendices A.3 and A.4) we identified ourselves as researchers

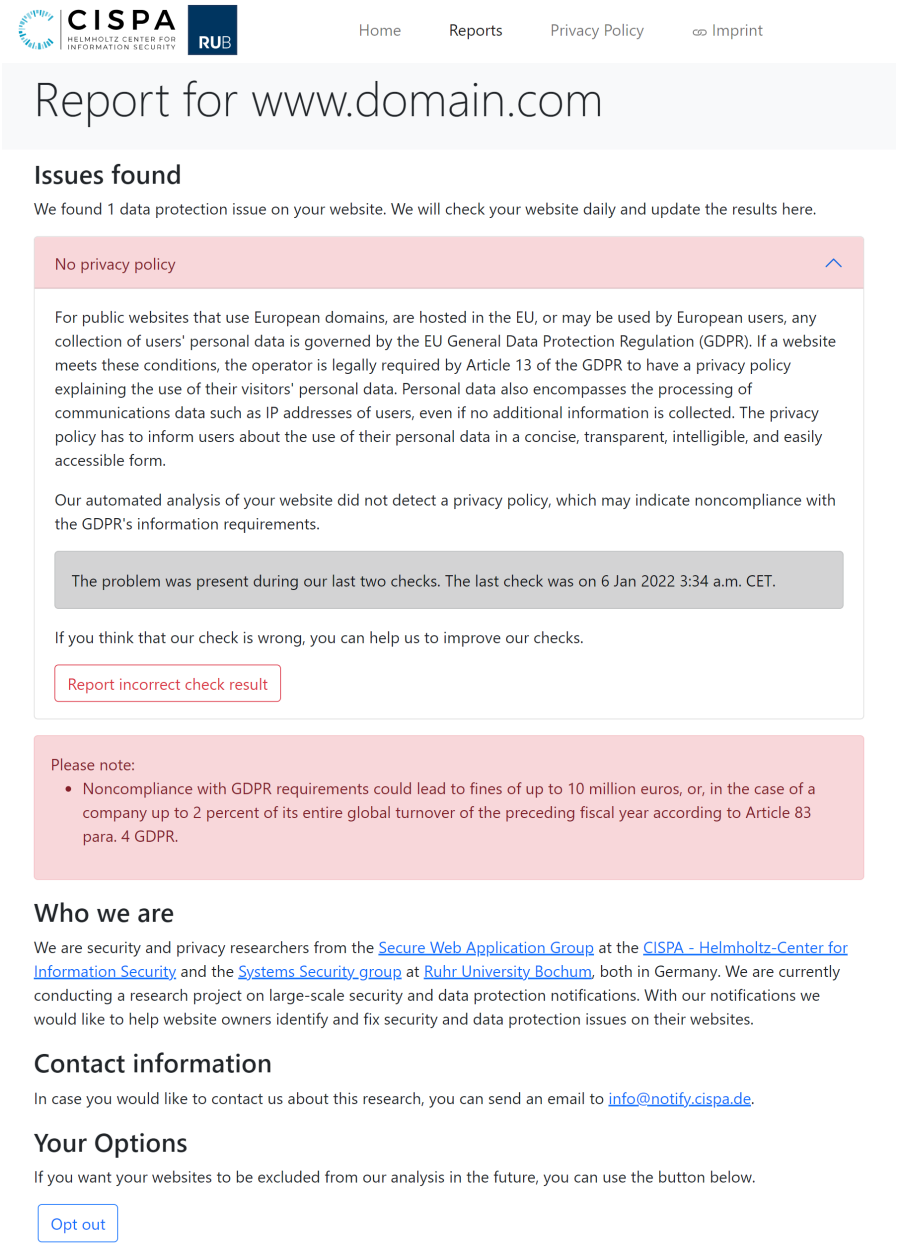


Figure 6.1: An example report for a website in the *Warning* condition (see Section 6.4.1) with a missing privacy policy. Notification recipients could access it via the link "web interface at `https://notify.cispa.de/reports/www.domain.com/report-[UNIQUE_ID]`" in the email. The color of the accordion menu for each detected issue changed to display its current status: red for not detected as fixed, yellow for detected as fixed once within the last two days, and green for detected as fixed for at least two days. The red "Please note" box was only shown for websites in the *Warning* condition and displayed the corresponding GDPR and/or ePrivacy and/or Git warning message(s).

and explained the purpose and scope of our checks and the whole research project. For the daily checks, we set the user agent of our OpenWPM crawler to `CISPA Web Analyzer (https://notify.cispa.de)` to point operators of the checked websites to our study website. Front office and IT staff at CISPA were briefed about the study, preparing them for operators potentially asking about the legitimacy of the study. Notified websites could use the opt-out functionality on the website with their report (see Figure 6.1) or send an email to be excluded from future checks. Web report accesses were collected in the report web application and disclosed in its privacy policy, drafted by our expert in data protection law. At the end of the study, we sent debriefing emails to still affected websites in the *Control* group (see Section 6.4.1), informing them about the detected security or privacy issues and our study.

Handling of reported false positives

To ensure that the privacy issues we notified websites about were universally acknowledged violations of privacy law, the study was conducted under the supervision of a legal expert with extensive knowledge of EU data protection law. In addition, we performed multiple rounds of manual verification of check results to make sure they produced as few false positive cases as possible.

Still, sending email notifications for complex privacy issues at scale meant that we inevitably reached out to some domains that were false positive cases for a privacy issue. When notification recipients approached us with reports of (presumed) false positives, we performed a timely manual inspection of their website and responded with the result to minimize recipients' time of uncertainty about the issue. We also routinely checked for false positives on domains whose operators had contacted us with an unrelated question. In total, 75 out of 414 email conversations with recipients of privacy notifications concerned (presumed) false positives. On 33 of these 414 domains we manually found true false positives (i. e., in 8 % of privacy conversations). The rest were presumed false positives, reasons for which we explore in Section 6.6.4.3. We assume that most recipients of a notification caused by a false positive contacted us before investing a significant amount of time in the issue. Nearly 50 % of true false positive cases could also be quickly identified by non-experts, such as websites actually having a privacy policy or not targeting people in the EU. Thus, we believe that we did not cause undue burden on notification recipients and the benefit for the other notified operators outweighed the potential cost for the few true false positive cases.

6.3.5 *Limitations*

Accuracy of checks

We defined technical constraints for privacy issues in collaboration with a legal expert, but since there are multiple steps involved that relied on external sources (e. g., for geolocation of IP addresses in the EU or classification of third-party requests), we can only aim to minimize, but not eliminate false positives.

Our study design also did not focus on avoidance of false negatives, so we likely missed many websites that in fact did have privacy issues. For example, if a site provided a link to a privacy policy, we did not further investigate if that page actually contained the necessary information mandated by privacy law.

Sample bias

Due to use in a pre-study, we removed `.de` domains affected by the *Git* issue from the set of domains. However, we do not believe this had any effect on our findings. We openly communicated that the notification process was part of a scientific study, possibly prompting fixes that would not have taken place otherwise due to the observer effect [168].

6.4 MEASURED NOTIFICATION RESULTS

Evaluating our measurements first requires us to describe the final study parameters we used when launching the notification campaign – domain set, experimental conditions, and notification schedule. After that, we present our results regarding website reachability, web interface usage, remediation rates, and the influence of warnings.

6.4.1 *Final Study Parameters*

NOTIFIED DOMAINS From the 388,825 domains initially considered, only 190,491 were still problematic when we started to send out notifications on October 20, 2021. The remainder was either fixed without our notification or could no longer be reached. To test our infrastructure, we sent out emails to 19,142 domains, which we removed from further consideration in our study. In this beta test, we noticed and fixed some minor issues. The full notification campaign started on November 3, 2021 and considered all 159,856 domains which were still flagged as problematic on November 1.

Control, Warning, and No Warning conditions

STUDY CONDITIONS Beyond the main focus of this study we wanted to explore whether warning about potential consequences of persisting issues (e. g., fines under the GDPR or national laws implementing the ePrivacy Directive), so far only investigated in a context limited in scope [169], had any effect on notification success. For this, we divided the domain set into three groups, each of which we assigned one out of three study conditions: a *Control* group (20 %); a group which received a *Warning* (40 %) about potential fines; and one that only received information about the issues but *No Warning* (40 %). The latter two are the experimental conditions. The concrete wording of the warnings can be found in the example notification in Appendix A.3.

Email type

In our analyses we also differentiate domains by *email type*, i. e., whether we contacted them via a *Parsed* or a *Generic* email address. We do not consider these "true" study conditions, as we did not have any influence on whether we were able to find an email address on a website.

Initial notifications, reminders, and debriefing

NOTIFICATION SCHEDULE We monitored the websites in the above data set over the course of two months, November and December 2021. Between November 3 and 5, 2021, we sent *initial notifications* to 297,506 email addresses associated with the 127,172 domains in the *Warning* and *No Warning* groups that still had the originally detected security and/or privacy issues as of November 1. Between November 20 and 22, 2021 we sent *reminder* emails to websites on which we could still detect the initial issue(s) as of November 18. Excluding bounced emails, domains for which the report had been viewed, and opt-outs,

Table 6.1: Number of domains (n) and prevalence of issues within each study condition and email type group. As the websites in our data set might have multiple issues, issue counts sum up to more than the total number of domains in each condition or group.

Condition / Group	n	*No Privacy Policy*	*No Consent*	*Before Consent*	*No HTTPS*	*Git*
Control	31,863	14,472	5,451	5,293	3,827	8,480
Warning	63,596	28,674	10,994	10,790	7,803	16,626
No Warning	63,576	28,750	10,832	10,700	7,654	16,816
Parsed	63,675	25,896	13,281	15,987	8,271	11,224
Generic	95,360	46,000	13,996	10,796	11,013	30,698
Total	159,035	71,896	27,277	26,783	19,284	41,922

reminders for 62,835 domains were sent to 98,079 addresses. We did not filter report visits for automated accesses, e. g., by URL scanners in email verification systems, but did not see any spikes in access rates in early November that would have been indicative of many automated accesses. Following the recommendation by prior work [168], we sent *debriefing* emails to 67,194 addresses for the 28,724 domains in the *Control* group on December 20, about seven weeks after the notification of experimental groups. The message text was identical to the initial notification for the *No Warning* group, including the links to the info website, report, and survey. Overall, we contacted 47,574 domains in the *Parsed* group via one email address, 8,122 via two, and 7,189 via three. All 93,832 domains in the *Generic* group were emailed via three generic addresses (see Section 6.3.3.1).

Final domain set

OPT-OUTS & FINAL DOMAIN SET Our initial set of notified domains comprised 159,856 domains. We received a total of 497 opt-out requests, 466 via the web interface and 31 by email. We also excluded 44 domains we had identified as false positives during email conversations (see Section 6.5). After removing another 280 domains that had turned out to resolve to a domain parking service, we were left with a final data set of 159,035 domains. Table 6.1 shows these domains by study condition, email type, and prevalence of identified issues.

6.4.2 *Reachability*

As expected with generic and automatically extracted email addresses, not all emails reached their recipients. Only for 70,542 (55.5 %) of initially notified domains at least one email was successfully delivered, which we assumed if our mail server was able to hand over the email to the recipient's mail server. While such a successful handover does not mean that an email will reach the recipient's mailbox, this metric still provides an upper bound for the notification delivery rate. The difference between *Parsed* and *Generic* email addresses was quite

high: While 87.8 % of initial notifications to *Parsed* addresses were successfully delivered, this was only the case for 33.8 % of emails to *Generic* addresses.

6.4.3 *Web Interface Usage Statistics*

Report views and opt-outs

As described in Section 6.3.3 and Appendix A.4, each notification email contained a link to our study website with more information about the project and a personalized report for each checked domain that also let participants report false positives or opt out of the study. Over the course of the study, 5,731 reports were viewed, 259 false positive checks were reported, and for 466 domains the web interface was used to opt out of daily checks. The number of opt-outs differed between issues: While 0.4 % (483) of the domains with a privacy issue opted out of our checks, only 0.1 % (58) of the domains with the *Git* issue requested their exclusion via the web interface. 75.2 % of report views occurred within 24 hours, indicating that the vast majority of recipients either reacted promptly or not at all.

6.4.4 *Remediation Rates*

To determine if our notifications had any measurable effect, we performed daily checks of the monitored websites over the course of two months.

6.4.4.1 *Sliding Window Approach*

To avoid domains incorrectly flagged as fixed because of one-off checker timeouts or page maintenance, we implemented a sliding window approach to determine if an issue persisted: For a given day t_i, we considered a domain d to be `problematic` if at any point in time within a 7-day window (t_i, t_{i+6}) our checker had identified the reported issue to still be present on the site. This 7-day window allowed us to obtain at least one real measurement result (i. e., a true / false evaluation of the checked issue, not a returned error or a missing data point) for 98.5 % of evaluated windows, resulting in a robust check.

6.4.4.2 *Evolution of Problematic Domains Over Time*

Figure 6.2 (a)–(e) shows for each issue, experimental condition, and email group the percentage of domains considered `problematic` at a given point in time. We investigated the persistence of issues more closely at four distinct points in time: one and two weeks after both initial notifications and reminders. The respective rates of `problematic` domains are shown in the % columns of Table 6.2.

Overall, we observe for a given privacy issue a similar downward trend in `problematic` domains across study conditions and email groups, with them mostly differing by only between 0–1 percentage points, though the *Control* group behaves as expected in yielding the highest rates of persisting issues. *Git* rates also follow this pattern but yield slightly higher differences to *Control*, in the dimension of 1–2 percentage points. This lack of a larger measurable effect is a direct result of our notifications' low delivery rates: With many domains in the experimental conditions never receiving a notification email, this subset is

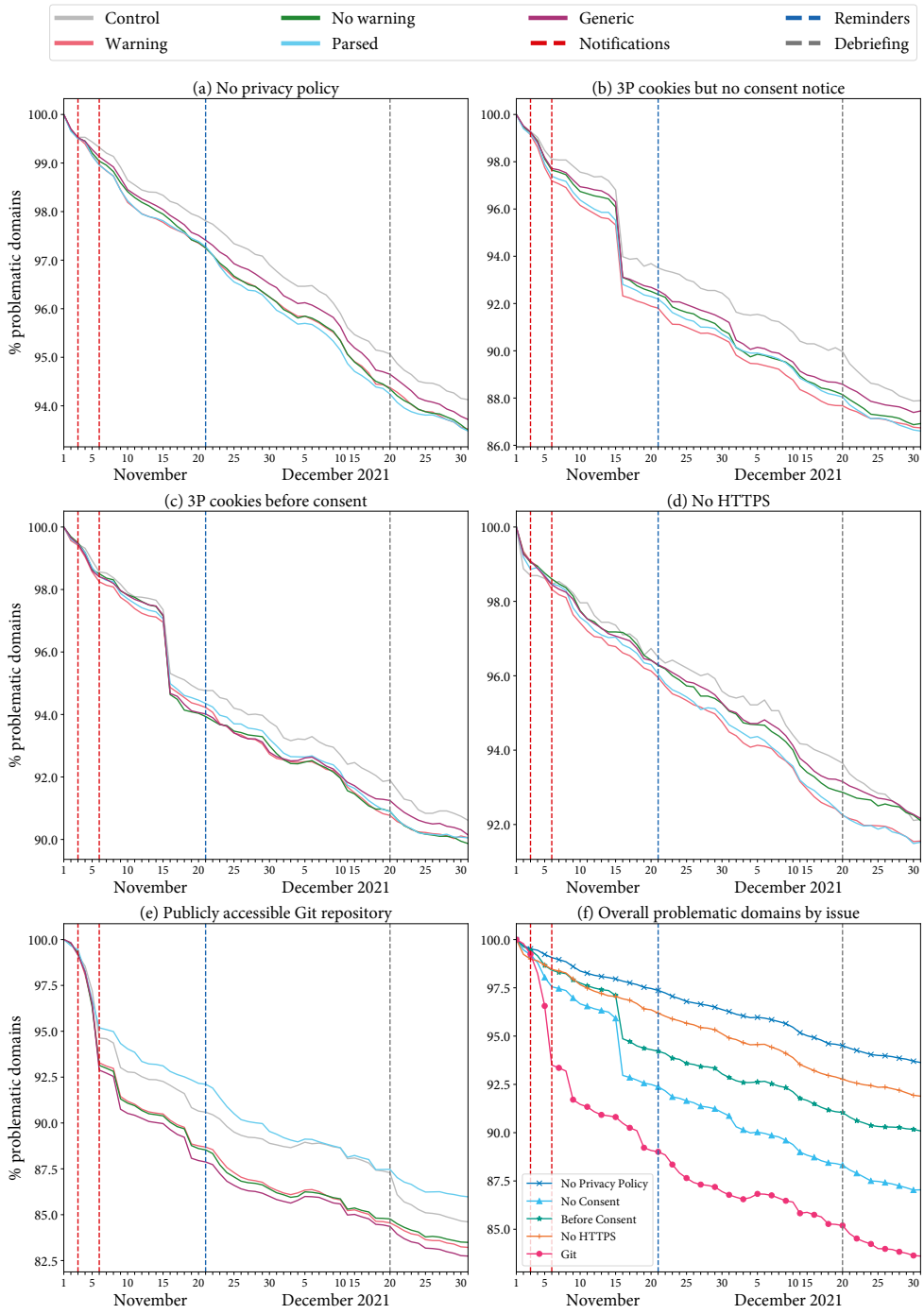


Figure 6.2: (a)–(e): Percentage of domains considered `problematic` according to our sliding window evaluation, by issue and study condition / email type. (f): `Problematic` domains by issue across all study conditions, including the *Control* group. Vertical lines of the same color demarcate the periods or points in time when notification emails were sent. There are no data points for December 11–13 due to necessary hardware maintenance.

expected to behave similarly to the *Control* group, exhibiting the same rates of natural decay of issues.

The "Twitter drop"

For the two issues related to the use of third-party cookies, *No Consent* and *Before Consent*, the number of `problematic` domains steeply dropped between November 15 and 16, 2021. Inspecting the affected domains, we found the cause to be a Twitter cookie named `lang`, originating from `cdn.syndication.twimg.com`, that was no longer present from November 16. This coincided with major platform updates at Twitter, including migration to Twitter APIv2[3].

Aggregated remediation rates

Figure 6.2 (f) compares the evolution of `problematic` domains by issue aggregated across all study conditions. Looking at that figure and the rates in Table 6.2, we observe that if it were not for the Twitter drop, there would be a consistent difference of about 5 % between the plot for the *Git* security issue and those for the privacy issues. This suggests that either websites were more inclined to fix security vulnerabilities or privacy issues required more time to address. Given the two-month period of our experiments, we could not conclusively figure out the exact reason. With slightly higher effects (diff column in Table 6.2), *No Consent* appeared to be the privacy issue easiest to remediate.

6.4.4.3 *Statistical Significance*

Fisher's exact tests

We also investigated the significance of differences in remediation rates at the aforementioned four points in time. We conducted Fisher's exact tests [84, 85] for the null hypothesis that the number of `problematic` vs. no longer `problematic` websites in the experimental conditions (*Warning* and *No Warning*) does not significantly differ from the *Control* group, and in an identical fashion for email types. Table 6.2 in its p columns shows the results for all four dates. We used the Holm-Bonferroni method [115] to correct for multiple testing over time.

While for November 10 the null hypothesis cannot be rejected for *Before Consent* and *No HTTPS* regardless of the presence of warnings, there is a significant difference in the distribution of `problematic` domains for *No Consent* and *Git* in both *Warning* and *No Warning* conditions.

These observations largely also hold true over time. The only cases where differences between *Control* and the experimental groups emerged later was *No Privacy Policy*, for which the *No Warning* condition did not lead to rejection of the null hypothesis on November 10 (while it did on any other date and across all dates for the *Warning* case), and *No HTTPS*, which only yielded significant differences to the *Control* group for the *Warning* condition on November 28, a week after the reminder. This could confirm an earlier security notification study that found a limited effect of reminders with web operators [260], but differences between issues suggest that some privacy issues take a longer time to be addressed. For email type, we observe similar tendencies over time. Across issues, *Parsed* email addresses more frequently resulted in statistically significant differences in fix rates compared to the *Control* group, except for the

3 See `https://developer.twitter.com/en/updates/changelog` under "November 15th, 2021."

Table 6.2: Percentages of websites still considered `problematic` with regard to each specific issue according to our sliding window evaluation (see Section 6.4.4.1) on the respective date in 2021, by study condition and email type. % indicates the percentage of still `problematic` domains, diff the difference in percentage points to the *Control* group, and p the p-values for Fisher's exact tests ($\alpha = 0.05$) compared to *Control*. Numbers in *italics* indicate values still significant after Holm-Bonferroni correction, while those in (brackets) indicate no longer significant values.

	Nov 10			Nov 17			Nov 28			Dec 5		
	%	diff	p	%	diff	p	%	diff	p	%	diff	p
No Privacy Policy												
Warning	98.19	-0.46	*0.0004*	97.61	-0.55	*0.0002*	96.46	-0.66	*0.0003*	95.85	-0.62	*0.0018*
No Warning	98.41	-0.24	0.0601	97.69	-0.47	*0.0015*	96.46	-0.66	*0.0003*	95.85	-0.62	*0.0018*
Parsed	98.22	-0.42	*0.0000*	97.63	-0.54	*0.0000*	96.37	-0.75	*0.0000*	95.70	-0.76	*0.0000*
Generic	98.45	-0.19	0.0830	97.83	-0.33	*0.0057*	96.72	-0.40	*0.0064*	96.12	-0.34	(0.0381)
Control	98.65	–	–	98.16	–	–	97.12	–	–	96.46	–	–
No Consent												
Warning	96.15	-1.41	*0.0000*	92.23	-1.64	*0.0001*	90.75	-1.80	*0.0001*	89.45	-2.09	*0.0000*
No Warning	96.74	-0.82	*0.0039*	92.98	-0.89	*0.0337*	91.28	-1.28	*0.0056*	89.85	-1.69	*0.0005*
Parsed	96.38	-1.18	*0.0000*	92.68	-1.19	*0.0003*	90.99	-1.56	*0.0001*	89.90	-1.64	*0.0001*
Generic	96.94	-0.62	*0.0138*	93.03	-0.85	*0.0166*	91.63	-0.93	*0.0201*	90.15	-1.40	*0.0009*
Control	97.56	–	–	93.87	–	–	92.55	–	–	91.54	–	–
Before Consent												
Warning	97.59	-0.31	0.2194	94.71	-0.49	0.1959	93.20	-0.81	0.0529	92.49	-0.71	0.1064
No Warning	97.83	-0.07	0.8166	94.49	-0.72	0.0602	93.32	-0.69	0.0939	92.48	-0.72	0.0993
Parsed	97.70	-0.20	0.3509	94.81	-0.39	0.1504	93.53	-0.48	0.1035	92.64	-0.56	0.0847
Generic	97.81	-0.09	0.6328	94.58	-0.62	0.0658	93.22	-0.79	(0.0406)	92.61	-0.59	0.1381
Control	97.90	–	–	95.20	–	–	94.01	–	–	93.20	–	–
No HTTPS												
Warning	97.42	-0.54	0.0815	96.54	-0.59	0.0962	95.05	-1.00	*0.0164*	94.13	-1.09	(0.0155)
No Warning	97.75	-0.21	0.4980	97.05	-0.08	0.8602	95.45	-0.60	0.1452	94.68	-0.54	0.2263
Parsed	97.57	-0.39	0.0838	96.75	-0.38	0.2229	95.14	-0.91	*0.0081*	94.37	-0.85	*0.0143*
Generic	97.73	-0.23	0.4293	96.94	-0.19	0.4649	95.61	-0.44	0.1681	94.72	-0.50	0.1904
Control	97.96	–	–	97.13	–	–	96.05	–	–	95.22	–	–
Git												
Warning	91.18	-1.60	*0.0000*	89.91	-1.91	*0.0000*	86.85	-2.35	*0.0000*	86.34	-2.61	*0.0000*
No Warning	91.08	-1.70	*0.0000*	89.80	-2.01	*0.0000*	86.71	-2.49	*0.0000*	86.25	-2.71	*0.0000*
Parsed	94.04	1.26	0.1438	92.64	0.82	0.8899	90.01	0.81	0.7519	89.12	0.17	0.0961
Generic	90.52	-2.26	*0.0000*	89.38	-2.44	*0.0000*	86.27	-2.93	*0.0000*	85.99	-2.96	*0.0000*
Control	92.78	–	–	91.82	–	–	89.20	–	–	88.95	–	–

surprising case of *Git*, where the effect of *Parsed* and *Generic* email addresses was reversed.

6.4.4.4 *Remediation by Website Popularity*

Websites' willingness to remediate security and privacy shortcomings may depend on available human and monetary resources and how well a site is maintained in general. Hence, we investigated if issues are less likely to persist on more popular websites. To obtain popularity metrics for as many websites as possible, we queried the Google Chrome User Experience (CrUX) data set [99] via BigQuery as described by Durumeric [57] for the full domain rankings from November 2021, when we sent out notifications. For domains listed twice due to CrUX differentiating between HTTP and HTTPS, we used the higher rank. This yielded an overlap between the CrUX data (8,733,078 origins, 8,567,511 domains) and the sites we monitored (159,035) of 21,592 domains, 7 of which were ranked top 1,000 by CrUX, 43 top 10,000, 432 top 100,000, 3,800 top 1 million and 17,721 top 10 million. Due to the low number of domains per popularity bin we focused on comparing remediation rates between CrUX-ranked domains (n = 21,592) and unranked ones (n = 137,443), i. e., the long tail.

Contrary to our expectations, for most issues and points in time we found the rates of `problematic` domains for CrUX-ranked domains to exceed those for unranked ones by 0–1 percentage points. For *Git*, this was even more pronounced, with differences mostly between 2–3 percentage points, except for the *Git–Parsed* combination, which followed the overall 0–1 % pattern. We presume this difference to be mainly due to issues "fixing" themselves naturally, with unranked websites being less reliable to reach and more likely to be taken down permanently. Breaking this pattern, for *No Privacy Policy* notifications to *Parsed* email addresses, CrUX domains consistently exhibited lower rates of `problematic` checks, though the difference also mostly lay between 0–1 percentage points.

6.4.4.5 *Influence of Warnings*

Logistic regression analysis

For more insights into the effect of the presence of warnings on fix rates, we took a closer look at the set of domains for which at least one email could be successfully delivered according to our earlier definition, i. e., handed over to the next mail server. To determine the influence of warnings on fix rates, we computed logistic regression models [178, 192] for each issue and four different points in time: one and two weeks after the start of sending initial notifications and reminders, respectively. Table 6.3 shows the regression models for all issues and experimental conditions (*Warning* / *No Warning*) relative to the *Control* group. For *Git*, *No Privacy Policy*, and *No Consent*, we observe a statistically significant influence of the *Warning* and *No Warning* conditions on fix rates compared to the *Control* group, while the models do not show such influence for the *Before Consent* and *No HTTPS* issues. Still, even in the case where *Warning* and *No Warning* performed significantly better than the *Control* group, we could not observe any difference between the estimates for these two conditions that did not fall within the standard error. This highlights that while receiving

Table 6.3: Logistic regression models for the remediation of each issue. Figures without brackets denote the estimates and figures in brackets the standard error. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

	No Privacy Policy				*No Consent*			
	Nov 10	Nov 17	Nov 28	Dec 5	Nov 10	Nov 17	Nov 28	Dec 5
Intercept	-5.81 ***	-5.44 ***	-5.12 ***	-4.89 ***	-2.48 ***	-2.14 ***	-2.01 ***	-1.94 ***
	(0.15)	(0.13)	(0.11)	(0.10)	(0.05)	(0.04)	(0.04)	(0.04)
No Warning	1.02 ***	0.93 ***	1.03 ***	0.90 ***	0.25 ***	0.18 **	0.25 ***	0.15 **
	(0.18)	(0.15)	(0.13)	(0.12)	(0.07)	(0.06)	(0.06)	(0.06)
Warning	1.24 ***	1.12 ***	1.23 ***	1.10 ***	0.31 ***	0.23 ***	0.27 ***	0.17 **
	(0.17)	(0.15)	(0.12)	(0.11)	(0.07)	(0.06)	(0.06)	(0.06)
AIC	3,684.49	4,654.99	6,407.46	6,873.09	11,089.59	13,312.77	14,516.55	14,165.17
BIC	3,710.53	4,681.06	6,433.49	6,899.06	11,112.98	13,336.19	14,539.95	14,188.50
Log Likelihood	-1,839.24	-2,324.50	-3,200.73	-3,433.55	-5,541.79	-6,653.38	-7,255.27	-7,079.59
Deviance	3,678.49	4,648.99	6,401.46	6,867.09	11,083.59	13,306.77	14,510.55	14,159.17
Num. obs.	43,512	43,809	43,453	42,450	17,978	18,166	18,027	17,608
	Before Consent				*No HTTPS*			
	Nov 10	Nov 17	Nov 28	Dec 5	Nov 10	Nov 17	Nov 28	Dec 5
Intercept	-2.51 ***	-2.23 ***	-2.17 ***	-2.09 ***	-2.53 ***	-2.40 ***	-2.34 ***	-2.31 ***
	(0.05)	(0.05)	(0.05)	(0.05)	(0.06)	(0.06)	(0.06)	(0.06)
No Warning	0.03	0.08	0.16 **	0.05	0.11	-0.04	0.07	-0.06
	(0.07)	(0.06)	(0.06)	(0.06)	(0.08)	(0.08)	(0.08)	(0.08)
Warning	0.04	0.08	0.12 *	0.07	0.17 *	-0.03	0.12	0.12
	(0.07)	(0.06)	(0.06)	(0.06)	(0.08)	(0.08)	(0.08)	(0.08)
AIC	10,585.40	13,030.21	13,795.53	13,546.50	6,876.26	6,956.61	7,550.38	7,341.52
BIC	10,609.04	13,053.88	13,819.19	13,570.07	6,898.50	6,978.86	7,572.60	7,363.68
Log Likelihood	-5,289.70	-6,512.10	-6,894.77	-6,770.25	-3,435.13	-3,475.31	-3,772.19	-3,667.76
Deviance	10,579.40	13,024.21	13,789.53	13,540.50	6,870.26	6,950.61	7,544.38	7,335.52
Num. obs.	19,490	19,739	19,625	19,077	12,226	12,304	12,174	11,915

	Git			
	Nov 10	Nov 17	Nov 28	Dec 5
Intercept	-2.68 ***	-2.54 ***	-2.31 ***	-2.22 ***
	(0.04)	(0.04)	(0.04)	(0.04)
No Warning	0.20 **	0.25 ***	0.32 ***	0.31 ***
	(0.06)	(0.06)	(0.05)	(0.05)
Warning	0.17 **	0.19 **	0.28 ***	0.28 ***
	(0.06)	(0.06)	(0.05)	(0.05)
AIC	12,232.64	13,602.02	16,084.50	16,927.29
BIC	12,256.86	13,626.23	16,108.70	16,951.48
Log Likelihood	-6,113.32	-6,798.01	-8,039.25	-8,460.64
Deviance	12,226.64	13,596.02	16,078.50	16,921.29
Num. obs.	23,664	23,654	23,538	23,498

a notification significantly improved remediation, the presence or absence of a warning did not. This contrasts with findings of prior work [169] that warnings did play a role in remediation success (albeit limited to German websites).

6.5 GATHERING RECIPIENT FEEDBACK

In order to gain further insights into the measured changes (or lack thereof), we leveraged two channels of communication with notification recipients to gather feedback: an online survey we linked in each notification message and email communication with recipients.

6.5.1 *Survey*

Surveys are a frequently employed tool in notification studies to learn about recipients' perception of the received notification and the underlying security or privacy issue(s) [34, 59, 157, 260, 299]. For consistency and to make sure participants received the survey invitation when the decision how to react to the notification was still fresh in their minds, we included the survey link in all emails we sent, i. e., the initial notification and reminder for domains in the experimental conditions (*Warning* and *No Warning*) and the debriefing message for the *Control* group.

Questionnaire overview

The survey was implemented using a LimeSurvey instance hosted at Ruhr University Bochum. The questionnaire first asked participants to assess the correctness of our checks (Q1), about prior awareness of the detected issue(s) (Q2), and plans to address them (Q3–4). Participants with privacy issues on their website were asked about the applicability of the GDPR (Q5–6), past changes to their website due to privacy legislation (Q7–8), and the influence of GDPR-mandated fines (Q9–10). Next, we asked all participants what type of support they would find useful to fix the issue(s) (Q11). We asked for participants' role(s) with regard to the website (Q12) to determine if we had reached a person with a suitable background to address the issue(s). The survey concluded with the opportunity to provide general feedback about our study (Q13). The full questionnaire can be found in Appendix A.5.

Analysis of open-ended survey answers

One of the researchers involved in this study conducted a manual thematic analysis on open-ended survey answers to inductively identify common themes and sentiments. We categorized the answers via labels informed by the survey questions and additional themes found in the answers. Survey responses are subject to self-selection bias, which includes people with a strong (negative) experience with our notification process being more likely to provide feedback.

6.5.2 *Email Communication*

Sending automated emails at large scale inevitably results in large volumes of incoming mail, including automated responses from ticketing systems, delivery status notifications, and "out of office" messages. To support participants with fixing identified issues and obtain more information how our notifications were perceived, we focused on incoming responses composed by humans.

Guidelines for communication

We answered emails in German or English, depending on the language used by the sender, and if requested, we also sent a German translation of our notification message. When asked for advice, we only referred to third-party resources that either had been published by the hosting company of a website, by a data protection authority in the website operator's country, or by the vendors of third-party software already used by the website. Upon request we provided the names of third-party cookies or URLs to Git repositories that had triggered a `problematic` check result.

Categorizing email conversations

The researcher who had signed the notifications answered incoming emails from notification recipients according to these guidelines and categorized them on the conversation level by identifying recurring themes. After about 50 conversations we found the topics to have reached saturation. From the resulting list we removed very rare codes, refined the definitions of the emerged categories, and added examples as well as counterexamples. Our final codebook (see Appendix A.6) comprised 23 codes in the following six categories, with the number of codes per category in parentheses:

- *Sentiment* (3 codes): Expressions of gratitude and positive or negative sentiments towards our project.
- *More information* (8): Requests for more information; for example, about our checks, the cookie or Git URL that had caused the website to be flagged, about our research project, and if our checks still detected an issue after changes had been made to the website.
- *Performed actions* (3): Status reports from notification recipients, including already fixed issues, plans to fix them in the future, or the decision to forward the notification email to the responsible people.
- *Correctness* (4): presumed false positives, including the website being outside EU jurisdiction or purportedly not processing visitors' personal information.
- *Language* (2): Sentiments concerning language, such as translation requests or consternation about the emails being in English instead of the language spoken in our institutions' or the recipient's country.
- *Other* (3): Including sentiments that the notification could be spam, verification requests to other points of contact at our institution, and opt-out requests.

The remaining emails were single-coded by two coders, with uncertainties resolved via discussion. Each conversation could be assigned an arbitrary number of codes. As we had only passed study parameters to the survey and no unique identifiers, we could not determine potential overlap between survey participants and email respondents, so some individually reported sentiments from these analyses may originate from the same person or domain.

Table 6.4: Notification survey participant sample (n = 212).

	n	%		n	%
Warning?			*Issues*		
Warning	95	44.8	No Privacy Policy	30	14.2
No Warning	117	55.2	No Consent	22	10.4
Email type			Before Consent	27	12.7
Parsed	105	49.5	No HTTPS	4	1.9
Generic	107	50.5	Git	140	66.0
Notification Round			Privacy only	72	33.0
Initial	149	70.3	Privacy & Git	2	0.9
Reminder	40	18.9	Git only	138	65.1
Debriefing	23	10.9			

6.6 RECIPIENT FEEDBACK

Our analysis of recipient feedback from email conversations and the survey identified largely overlapping topics. We briefly characterize the gathered data sets before investigating these recurring themes.

6.6.1 *Overview of Survey and Email Responses*

Response rate

SURVEY PARTICIPANTS Overall, the survey link was clicked 1,890 times. 1,556 people only accessed the welcome page, 121 provided partial responses, and 213 completed the survey. We discarded all incomplete responses, plus one response for which the survey parameters had not been passed and questions based on specific detected issues had not been displayed. This left us with 212 complete responses.

Sample characteristics

Table 6.4 shows the sample of participants who took the survey by experimental condition, email type, detected issues on the website, and notification round. While full responses were roughly equally distributed between the *Warning* and *No Warning* conditions, as well as email address type, most responses were collected through initial notifications (including the control group debriefing), as opposed to the reminder. This corresponds with earlier findings that notification recipients either tend to act upon the first received email or not at all [157]. Two thirds of survey participants had been notified because of an open Git repository, while privacy issues less frequently motivated people to take the survey. This hints at security notifications being either taken more seriously or at least leading to higher willingness of recipients to interact with us. Appendix A.5 provides an overview of all survey questions with response counts; the analysis in this section focuses on selected themes.

Email statistics

EMAIL COMMUNICATION We received a total of 760 emails in 621 conversations with the operators of notified domains. Contrary to the distribution for the survey feedback, 414 of these domains had been contacted because of a

privacy issue and 167 because of a publicly accessible Git repository. 19 emails could not be assigned to a domain due to a lack of provided information.

Verification through other channels

The majority of these emails (662) had been sent to the two email addresses designated for this study. In addition, we received 20 emails forwarded by the CISPA front office, and 85 had been sent to the institutional email address of the researcher who had signed the notification email. This was mainly done to verify if our notifications were legitimate. 7 emails had been sent to both the researcher and the project addresses. Appendix A.6 shows how often each code (see Section 6.5.2) was assigned to email conversations with domains with security and privacy issues.

6.6.2 *Who Did We Reach?*

Respondents' website-related roles

If emails were successfully delivered, they had significant impact on fix rates, so we wanted to understand who the recipients were. In the multiple-choice Q12 the majority of survey participants reported technical roles (developer and similar 57.1 %; administrator / operator 60.4 %), followed by roles related to the website's content (19.8 %) and product / project management (13.7 %). As for people in legal advisory roles, data protection officer ranked fifth 9.9 %), while the role as legal counsel was reported less often (2.8 %). The involvement of these two roles was equally distributed between *Git* and the privacy issue(s). While survey participants only represent a fraction of emailed websites, these numbers provide evidence that the people who ultimately felt incentivized to react to the notification mainly hold responsibility for a website's technical administration or content.

Referrals to other people

In email conversations we looked for how often the recipient referred the handling of the issue to another person (code: *notified*). This was the case in 15.9 % of privacy conversations and in 10.2 % of those about the *Git* issue. Explicitly mentioned people or entities for both types of issues included the IT department or webmaster, in the *Git* case the security team, and for privacy notifications a lawyer, the cookie plugin provider, or the marketing department.

6.6.3 *When Do Recipients (Plan to) Remediate?*

Beyond the daily website checks, we wanted to gain additional insights into notification recipients' remediation behavior and future plans to address the issue(s) (or not). For this, we analyzed recipients' feedback regarding awareness of the issues and their willingness to remediate them.

Awareness of issue(s) and (planned) changes

In survey Q2 we found that while most participants (72.6 %) reported to have been unaware of the issue(s) prior to receiving the notification, this number was higher for security (81.4 %) than for privacy issues (56.8 %). Participants' reported plans to make subsequent changes to the website (Q3) also differ: While overall 81.6 % planned to make changes, this applies to 90.7 % of participants notified of *Git* issues but only 64.9 % of people with privacy problems. Pairing these results from Q2 and Q3, it seems that privacy issues are more often knowingly ignored.

Email analysis revealed similar differences in remediation intentions. Privacy notification recipients mostly told us that the issue(s) would be handled in the

future (37.9 %, code: *will-handle*), while only 14.5 % stated that they had already been fixed. For the security notification, we saw the opposite: For *Git*, in 16.2 % of conversations the recipients stated that the issue would be handled in the future, while 44.9 % reported that the issue had already been fixed. As in the survey responses, this could either mean that privacy issues are more likely to deliberately be left unfixed – or that fixes simply take longer as they are more complex and may require the involvement of legal professionals, for example, to draft a privacy policy.

6.6.4 *Roadblocks to Notification Success*

Qualitative feedback from the survey and emails also allowed us to identify factors that hindered recipients from taking remediative action upon receiving our notifications.

6.6.4.1 *Language Barrier*

As we emailed domains from various countries in English, we may have contacted recipients in a language they do not understand, as first indicated by 3 survey participants in Q13 who found it hard to understand or assess the trustworthiness of an email written in English. More concrete evidence were the 17 translation requests we received via email. Most requests were for German, but some also for French and Czech.

6.6.4.2 *Notifications Perceived as Spam or Otherwise Malicious*

When asked for general feedback in survey Q13, 9 out of 76 participants (11.8 %) mentioned that they initially had been suspicious that the notification was spam or a scam attempt. To fight this impression, one participant suggested to point out the S/MIME signature in the email body, as "[s]pammers don't go out of their way to sign their emails from a public CA issued PEM certificate" (P1254[4]).

Verification requests

Similarly, email analysis found in 7.2 % of conversations about *Git*-notified domains and in 12.1 % of privacy-related correspondence that the recipient was not sure if the notification email had been sent with benign intentions (code: *unsure-scam*). A special case were emails asking if the notification email had really been sent by our institution. 4 such verification requests were sent to dedicated project email addresses, 35 to the institutional address of the author who had signed the notification emails, and another 14 were sent to CISPA's front office.

Notification language

The language of the notifications may also have contributed to this. 6 email respondents wondered why we, as researchers from German institutions, sent emails in English (code: *expected-german*); 3 of them stated they were not sure if our email was benign. This was also observed by Li et al. [157], who reported that emails the recipient expected to be in a different language (e. g., based on the sender's country of origin) were sometimes considered phishing or spam.

4 Participant IDs refer to the response number assigned by LimeSurvey.

6.6.4.3 *(Perceived) Incorrectness of Reports*

Reported (in)correctness rates

In survey Q1 the majority of participants answered that the report was correct, with rates highest for *Git* (87.9 %) and lowest for *No Consent*[5] (45.5 %). Correspondingly, reported incorrectness rates were lowest and highest for these two cases, with 5.7 % and 22.7 %, respectively. Uncertainty about the correctness of the report was highest for the *Before Consent* (29.6 %) and *No Consent* (18.2 %) cases. This hints at notification recipients often finding it difficult to determine which third-party cookies were present on their website and if they required user consent. While not a true false positive for the *Git* check, 3 survey participants notified about it replied in Q4 that they did not intend to make changes because the issue did not pose a security risk, as their Git repository did not contain any sensitive information, was not under their control or not accessible.

No sensitive data stored in Git

In emails we also received feedback about check results being false positives, for 18.1 % of conversations about a privacy issue but also for 6.0 % of conversations about publicly accessible Git repositories. 16 emails stated that no sensitive data was stored inside the Git repository, though 3 reported that they had still made the repository inaccessible.

Failed attempts to reproduce the issue

Manual checks revealed the majority of false-positive claims for the *Git* case to be due to failure to reproduce the issue. Many recipients of *Git* notifications tried to access `<domain>/.git/`, saw a "Forbidden" error page from their web server, and falsely assumed that this meant the Git repository was inaccessible, while in fact directory listing was forbidden. It is likely that they then stopped to further investigate the issue, leaving it unfixed.

6.6.4.4 *Perceived Inapplicability of Privacy Legislation*

One recurring theme in both the answers to multiple survey questions and email conversations about privacy notifications was recipients' perception that the privacy legislation in question did not apply to their website. In survey Q5 the 74 participants notified about a privacy issue were asked whether they thought the GDPR applied to their website. 63.5 % thought it did, while 20.3 % did not think so and 9.5 % were not sure.

Territorial scope

When asked in Q6 why they thought the GDPR did not apply, we received 14 replies. 6 answered that they were not located in the EU ("Because the UK is no longer in the EU" [P349]), illustrating unawareness of the GDPR's extraterritorial applicability, which extends to non-EU websites with EU visitors (see Section 2.2.1.1). Interestingly, several of these respondents were based in the UK, which, despite having left the EU in 2020, still has a verbatim copy of the GDPR in its national legislation, so the same legal requirements apply based on domestic UK law.

Material scope

Regarding the material scope of the GDPR, 7 participants claimed to not process any personal data, unaware that even temporary storage of IP addresses is considered processing of personal data under EU law (see Section 2.2.1.1).

5 We do not consider *No HTTPS* here, as only 4 survey participants had this issue.

6.6.4.5 *Privacy Indifference*

Another demotivating factor that only emerged for privacy notifications was a general disdain for privacy legislation and its requirements. We found such sentiments in the answers to multiple open-ended survey questions. Asked in Q4 why they did not intend to add a privacy policy to their website, one participant replied "because these rules are plain stupid!" (P1131), and in Q8 another refused to make any changes at all due to the GDPR: "does not matter – GDPR is sucks [sic]" (P671).

6.6.5 *Motivation for Remediation*

In the opposite direction, the survey also explored factors that motivated participants to (want to) take action, particularly awareness of fines mandated by the GDPR (see Section 2.2.1.6).

Awareness of fines

In Q9 more than half of the 74 survey participants with privacy notifications (38, 51.4 %) had already been aware of these fines before the notification. Another 9 (12.2 %) had learned about them via the email, with 7 participants in the *Warning* and 2 in the *No Warning* condition. Still unaware of fines were 20 participants (27.0 %), 8 from the *Warning* and 12 from the *No Warning* group. Half of the 8 warned but unaware participants had only seen ePrivacy warnings, but the other 4 had (also) been warned about GDPR-mandated fines. This shows that notifications have limited educational impact about privacy legislation and potential fines.

Impact of knowledge about fines

Q10 explored how knowledge of GDPR-mandated fines had influenced participants' decision to fix the detected issue. 13 participants (27.7 %) answered that they had not been influenced by the risks of fines, stating as their motivation that they "wanted to be responsible" (P45) or "believe[d] that GDPR and compliance with it [was] important" (P1505). Another 13 (27.7 %) explicitly acknowledged that fines were a motivating factor in their decision ("want to prevent paying fines" [P973]). 7 of them had received a warning notification and 6 an email without a warning. Though these answers may suffer from social desirability bias, this hints at fines for GDPR non-compliance being known and influencing fix rates, regardless of whether they were explicitly mentioned in the notification.

6.6.6 *How Can We Help Websites Fix Issues?*

The above findings leave the question how future research can help website owners fix identified privacy issues.

Survey Q11 asked what additional information recipients would have wished for to better understand and fix the notification issue(s). We received 129 open-ended responses, many of which expressed generic sentiments: 37.7 % found the notification helpful and the information to be sufficient.

More detailed notification messages

12.3 % would have appreciated more detailed guidelines or links to external resources on how to fix the issue(s). 16.5 % asked for additional documentation of our checks: 15 *Git*-notified participants suggested to add the URL for the repository in question, and 2 asked to include a check whether any sensitive

information was present in the repository. While the first suggestion is an easy fix for future notifications, the latter would require extensive resources and would raise ethical concerns. Regarding the use of cookies without consent, we were repeatedly asked to add to the notification the names of the third-party cookie(s) that had triggered the `problematic` flag, which is also feasible.

More information about checks

Classification of email conversations confirmed this. A major category were requests for more information: 8.0 % of the 414 conversations regarding domains with privacy issues asked about privacy checks, especially the name of the problematic cookie (5.3 %); and questions concerning *Git* checks (10.8 % of the 167 *Git* conversations) most often were interested in the URL of the publicly accessible repository (5.4 %). We also received more generic requests, such as what to do in general, if certain changes would make the website compliant with privacy law, or about our research project.

Addressing misconceptions

Addressing recipients' misconceptions about the often complex material and territorial scope of privacy legislation is more challenging for researchers. While future notification campaigns could preemptively include more detailed explanations on these questions, this could make the notification email too verbose, which could impact its effectiveness.

6.6.7 *How Were the Notifications Perceived?*

It should be best practice in security and privacy research to notify affected parties about potential security or privacy issues identified on their systems, but there is a risk of backlash that could harm individual researchers and the community.

Survey

Hence, in survey Q13 we asked for general feedback about our project and received 76 responses. Sentiments varied greatly between recipients of security and privacy notifications. 76.0 % of respondents notified about *Git* thanked us for the notification or voiced positive feedback ("This is an amazing project, please keep up the good work to make the internet a more secure place!" [P1675]), but only 50 % of respondents with privacy notifications did so ("Thank you for this hint! There are so much [sic] rules. For a little webmaster it's hard to know everything. It's really great to know there are some who help the little ones ;-)" [P840]). Negative sentiments were only expressed for privacy notifications ("the stated analysis is only 'may be' ... You have just wasted our time & energy" [P1685]). Repeated criticism included the privacy notifications being false positives, too threatening, unwanted, not sent in the participant's language, or sent with ill or monetary intentions. This confirms the sentiments reported in Section 6.6.4.

Email analysis

Email feedback contained similar differences in sentiment. Here people thanked us for the notification (code: *thanks*) in 74.9 % of conversations with recipients of security notifications, but only 56.0 % of privacy conversations. To distinguish between "thanks" and real enthusiasm for our project, we used the code *great-project*, assigned to 16.2 % of security and 2.9 % of privacy conversations. Correspondingly, the distribution of negative sentiments towards our notifications or project was reversed, assigned to 5.3 % of privacy- but only 1.8 % of security-related email conversations.

Overall, the feedback received via these two channels paints a consistent picture: Website operators tend to be more favorable towards notifications about security issues than those about privacy problems.

6.7 DISCUSSION

In our investigation of whether large-scale email notification campaigns can motivate websites to remediate complex privacy problems, we identified recommendations both for future research in web privacy as well as for the public entities tasked with the application and enforcement of privacy legislation.

6.7.1 *Privacy vs. Security Notifications*

We compared the effects of notifications about privacy issues with those about a security problem. Challenges faced by both types include how to reach the responsible parties, language barriers, and lack of a trustworthy messaging channel. These have already been identified by previous work on security notifications and continue to pose significant problems for any campaign that aims to reach people via email at scale, as it is hard for both computer systems and humans to differentiate automated emails sent with beneficial intent from those sent for malicious purposes. Specific to privacy notifications is the obstacle that many recipients are not aware that certain legal requirements apply to their website, either because of misconceptions regarding the territorial scope of privacy laws or the website's data processing operations. Future research in this area is encouraged to educate notification recipients about the applicability of privacy laws and provide concrete information why the respective law was deemed to apply to the recipient's website. This is especially important given our observation that the motivation to make changes due to privacy notifications appears to be extrinsic (awareness of fines) more often, while fixes of security issues tend to be more intrinsically motivated.

6.7.2 *Message Tone and Content*

Alleviate perceived threats

We found security and privacy notifications to be met with very different sentiments, which could be rooted in message content (mentioning of privacy laws) or tone (presence of warnings). Security notifications evoked more positive sentiments and fewer perceivd threats of legal action. To relieve recipients' anxiety and anger, we recommend researchers in future privacy notification studies to *explicitly* explain that they will not pursue any legal action against notification recipients.

Concrete information

Recipient feedback also indicated widespread problems to identify the third-party plugin or subpage that had triggered the placement of a third-party cookie before or without the visitor's consent. We recommend that future notification studies provide the necessary details to help recipients pinpoint the problem, in our case the Git URL or concrete third-party service or cookie and subpage that had triggered the detection mechanism. Links to concrete guidelines by data protection authorities or courts (e. g., that pre-ticked checkboxes do not consti-

tute informed consent [69]) could aid notification recipients in understanding the issue and why it applies to their website.

6.7.3 *Call for Guidelines and Standardization*

Our results demonstrate that it is generally feasible to identify privacy issues on websites independent of specific vendors or consent frameworks, with a prevalence of false positive cases in the low single digits in our set of manually verified websites. Still, these checks were designed to minimize the number of false positives for the purposes of this notification study and false negatives were not a concern. Consideration of false negatives may well be a requirement in other contexts, such as automated privacy audits designed to take the burden off DPAs in enforcing privacy legislation. Identifying vendor-agnostic privacy issues at scale with low numbers of both false positives and false negatives is still a significant challenge, given the differences in the implementation of, for example, privacy policies or consent to the processing of personal information. Regulators could aid privacy researchers – and themselves in enforcing privacy laws – by issuing more concrete guidelines how to implement requirements posed by privacy law, as in the example of the CCPA that requires a "Do Not Sell My Personal Information" link (see Section 2.2.4). On a more general level, the persistent challenge to enforce privacy laws online in the light of limited human and monetary resources requires long-term assistance through automated audits. The challenges associated with vendor-agnostic assessments could be alleviated via standardization of how privacy-related information is presented on the Web. Hence, web researchers and standardization committees are encouraged to create new and build upon existing proposals how to unify the presentation and user control of a service's data processing practices. To prevent any such standard from suffering the same fate as past web privacy mechanisms relying on voluntary adoption, such as DNT or P3P, regulators should make it mandatory for web operators and browser vendors to adhere to these standards.

6.7.4 *The Challenge of Reachability*

Email addresses extracted from websites

Our use of email addresses found on websites yielded promising results in terms of reachability: 87.8 % of initial notifications to *Parsed* email addresses were successfully delivered, while this was only the case for 33.8 % of emails to *Generic* addresses. While this does not guarantee that the correct person is reached and they act upon the notification, this approach can help overcome one of the obstacles in reaching the people who are responsible for fixing security or privacy issues on websites. It needs to be noted that this approach can possibly introduce bias into the data set of notified websites, as well-maintained websites are more likely to provide contact information, including an email address.

Standardized point of contact

The reachability problem also provides an opportunity for standardization: In the vein of `security.txt` [86, 87], a proposed standard to help web security researchers identify points of contact for vulnerabulity notifications, a file `privacy.txt` could serve this information for privacy-related issues – and also be used to communicate information of a website's data processing practices in

a standardized format or at least contain a link to its privacy policy. Until then, future privacy notification studies could also leverage contact information from `security.txt` to notify websites about privacy issues.

6.7.5 *The Future of Privacy Notifications*

Our results show limited success of our notification campaign, especially when weighing this outcome against the resources required to detect and notify websites about privacy issues at scale. Similarly, the improvements proposed above will be futile if recipients of notification emails do not trust the sender do not consider privacy violations a problem whose remediation is an urgent matter. Still, we believe that large-scale privacy notifications can be a valuable tool in improving web privacy, but they need to be accompanied by other measures to overcome these obstacles.

Joint campaigns with DPAs

To increase sender credibility and authority, researchers could cooperate with data protection authorities for future notification campaigns about issues related to data protection and privacy. The role of the DPA would be to provide sender credibility via their legitimization as a public authority and to accompany the campaign with information and enforcement capabilities to raise awareness for the issues. They could communicate to the general public the goal of the notifications, participating research institutions, investigated issues, guidelines on how to fix them, territorial and material applicability of the relevant laws, and possible consequences of non-compliance. A DPA informing about the latter is likely to be met with less adversity, as enforcing applicable privacy laws is their core task. Researchers, in turn, could supply the personal and technical resources that public authorities often lack and provide the notification infrastructure, expertise to more reliably detect compliance issues at scale, and (limited) support with fixing them.

Past campaigns show that this could be a promising approach. Some DPAs already have experience with privacy-related web measurements, such as the "Cookie Sweep" [13] carried out by multiple national DPAs in 2014 to inform EU institutions about websites' use of cookies and obtain first evidence for ePrivacy compliance. The notification campaign by privacy NGO noyb [197] (see Section 6.2.2) shows how external entities can support authorities in enforcing privacy legislation. Manual analyses limited the scope of these campaigns, but they both illustrate where DPAs and privacy researchers could benefit from each other to help enforce privacy legislation at scale. These multi-modal campaigns might not only have a broader effect, but also provide opportunities for further research, such as evaluating the usability of guidelines to fix privacy issues.

6.8 CONCLUSION

In this chapter, we conducted a large-scale email notification campaign to investigate if this approach is also viable to help websites fix more complex privacy issues like missing privacy policies and incorrectly implemented consent notices and to determine how they compare to notifications about security vulnerabilities. Though overall fix rates are higher for security than privacy issues and the latter show tendencies to be addressed at a later point in time, we

still find a statistically significant influence of our notifications on remediation rates. To overcome the problem of websites being hard to reach, we identify a promising approach in automatically extracting contact information from websites.

Qualitative feedback from email conversations with recipients and survey responses hints at website owners being less open towards notifications about privacy issues than a security vulnerability. Reasons include limited willingness to make changes for privacy compliance, widespread misconceptions about the applicability of privacy laws, and often greater necessary effort to identify and fix the problem. Even though warnings about potential implications such as fines do not increase remediation rates, they do at times incur anxiety and anger with recipients and corresponding backlash towards the senders. Future work is encouraged to explore if more specific information about the privacy problems and assurance of benign intent can yield more positive reactions and make email notifications a tool that can support large-scale privacy compliance.

Part IV

EPILOGUE: TOWARDS A MORE USABLY PRIVATE WEB

7

FUTURE WORK

In this work we have studied third-party web tracking under the GDPR from two perspectives: website visitors, who are subject to ubiquitous consent notices that only allow limited control of tracking through third-party services, and web developers and people in related roles who integrate these services into websites, often unaware of their privacy implications. Similarly, for future work aiming to improve website visitors' privacy with regard to third-party web tracking, there are research opportunities addressed at different entities involved in the process.

7.1 EASING THE BURDEN ON WEBSITE VISITORS

On the visitors' side, the existing situation – ubiquious consent notices that are often perceived as annoying and ineffective – could be improved by reducing the amount of consent decisions or tying them more closely to the associated data collection processes.

7.1.1 *Moving Consent to the Browser*

Since the study presented in Chapter 3, the market for online consent functionality has consolidated [114] towards vendors that support the mechanism we dubbed *centralized consent management* in Section 3.4.2.2 – IAB Europe's Transparency and Consent Framework [132], created and backed by the online advertising industry. This has given rise to ever more increasingly complex consent notices designed to maximize acceptance rates through use of dark patterns while providing the impression to comply with EU privacy laws. In addition to the formatting techniques we explored in Chapter 4, these include hiding available options to deny consent on a second, initially invisible layer and artificially introducing delays [39], with the goal to make the opt-out click path take multiple times as long as the "accept all" option [39, 114]. Paired with the fact that these notices were not always found to be correctly implemented to honor visitors' consent decision [174], this leaves users annoyed and fatigued with privacy-related decisions, taking the fastest route to get rid of the consent notice that blocks access to the website content: select the "accept all" option and rely on browser plugins to block unwanted cookies and advertising [89].

From a UX perspective the ideal solution would be not to display any consent notice at all but let website visitors specify their privacy preferences at a single point, for example, in the browser, which then communicates these preferences to websites [89]. Past efforts in this area like Do Not Track (DNT) [83] have failed due to websites' lack of interest and incentive to cooperate. DNT is also not capable of implementing purpose-based, specific consent as mandated

by the GDPR [89], but the standard could be extended to include signals that differentiate between specific predefined purposes [200].

Likely fueled by the unsatisfactory state of consent on the Web, conceptually similar approaches have recently seen renewed research interest. One of them, Global Privacy Control (GPC) [96], is already backed by multiple vendors of browsers and privacy-focused browser extensions, digital rights groups, and publishing houses. Users of supporting browsers or extensions can configure them to include a GPC signal in the headers of HTTP requests for all websites they visit or on a per-website basis.

What differentiates this mechanism from DNT is that the California Attorney General explicitly clarified that the GPC signal is considered a "Do Not Sell My Personal Information" request under the CCPA (see Section 2.2.4) and websites are legally obliged to comply with this request [257]. Thus, supporting websites currently interpret the signal as such and treat the visitor as if they had opted out using the website's own "Do Not Sell" mechanism. In the future GPC could be extended to support exercising privacy rights under other jurisdictions, including the GDPR – though it remains unclear to what extent a browser-based mechanism would be able to satisfy the GDPR's requirement that consent be *specific*, *informed*, and *unambiguous* [236]. Another recent proposal aiming to automate consent under the GDPR and the planned ePrivacy Regulation is Advanced Data Protection Control (ADPC), devised by a consortium of researchers from the University of Vienna and privacy NGO noyb [124]. If jurisdictions beyond the US state of California create similar legal obligations to obey privacy signals, a browser- or device-based approach could reach higher acceptance. Still, the actual effectiveness of such standards also relies on websites correctly implementing the conditional triggering of data collection processes through third parties, which, as we have seen in Chapter 3, is already often not the case with consent notices. Human et al. [123] compared ADPC, GPC, and earlier proposals for the automated communication of privacy preferences and identified legal, technical, human-centric, and other challenges such mechanisms need to overcome to gain widespread adoption.

7.1.2 *Putting Consent into Context*

In Chapter 4 we have shown that website visitors often have misconceptions about the workings of consent notices on websites, particularly the fear that the website could not be accessed or would not work properly if they denied consent to data collection. One possible reason could be a temporal and spatial disconnection between the consent prompt and the actual data collection process, as previous work has shown that, like privacy preferences in general, those regarding web tracking are highly context-dependent [182].

Hence, effective privacy controls should, if possible, be shown in the temporal and spatial context of the associated data practice [237, 238]. Recently some websites, particularly news websites in German-speaking countries, have started to use two-click mechanisms for different types of embedded third-party content, particularly from social media. As shown in Figure 2.1 (c) in Chapter 2, these consent mechanisms are built to be shown in place of the embedded third-party element, and if the visitor clicks to agree to the involved

data processing, the consent element is replaced with the remote content, triggering data transmission to the third party. This directly ties the consent action to the data processing action. We hypothesize that this contextualization makes it easier for website visitors to understand what data processing they are consenting to by clicking the button. By granting access to the surrounding content before consent, these mechanisms could also alleviate fear that it is the entire website that cannot be accessed or will not work properly if consent is denied. In contrast to blocking consent notices, this approach also allows website visitors to simply ignore the third-party content, such as an embedded social media feed, and avoid making yet another consent decision.

Future work could evaluate and compare users' perception of and their interactions with these contextualized consent mechanisms and compare them against cookie consent notices, either in a lab setting or in the field like in our study presented in Chapter 4. One thing to keep in mind is that this approach can only be applied to data collection processes that can be reasonably tied to a certain location on the website (contrast to, e. g., website analytics, which typically use tags embedded all over the website). It is also not conceivable that websites would use such mechanisms for third-party advertising, as in this case the first-party website relies on the ads to be seen. In addition, individual consent elements for different types of functionality can lead to an even higher number of consent decisions instead of bundling them into a consent notice. Future work could evaluate how many of these contextual consent decisions website visitors are willing to make, as there could well be a trade-off between increased understanding due to contextualization and increased negative sentiments if too many of these elements are shown.

7.2 ENCOURAGING "PRIVACY BY DESIGN AND BY DEFAULT" WITH WEBSITES

In Chapters 5 and 6 we studied the people who are ultimately responsible for the inclusion of third-party services into websites and found a widespread lack of awareness regarding the data collected through third-party services, including the use of cookies without or before consent. There is plenty of research opportunity in finding ways to encourage the use of less privacy-invasive technology on websites.

7.2.1 *Incentivizing Privacy in Web Development*

In Chapter 5 we have seen that the people working on websites frequently do not consider alternative integrations for a desired functionality or make privacy-friendly configurations because they often base their decision on ease of integration and familiarity with the chosen solution and tend to underestimate the amount of data collected by third-party services. In addition, the notification study in Chapter 6 found that people often struggle with pinpointing the use of third-party cookies without consent or understanding its privacy implications.

Coding experiments

The restricted formats of online surveys and asynchronous communication via email can only provide limited insights into the underlying reasons and opportunities to raise awareness and incentivize change. A more thorough un-

derstanding of the decision processes that lead to the integration of third-party services into websites could be obtained in a study design that is more interactive, such as interviews, or that more closely resembles participants' actual work environment. The latter can be achieved via controlled programming experiments [264], previously used to study various developer behaviors including security practices like passwort storage [2, 49, 188–190] or use of cryptographic libraries [1, 275]. Prior research exploring the security and privacy practices of developers has shown that priming for security [189] or privacy [265] can encourage more secure or privacy-friendly integrations or configurations. Future work could draw inspiration from these study designs and investigate the influence of priming for privacy on the selection and configuration of technology to integrate different types of functionality into websites. Such a study could not only investigate participants' existing awareness, knowledge, and resulting privacy practices in this largely unexplored domain but also determine factors that would increase adoption of more privacy-friendly technology.

7.2.2 *Alternative Website Business Models*

Any long-term improvement of websites' privacy practices needs to take into account their owners' legitimate business interests. In Section 2.1.3 we already established that much free Web content is monetized via online advertising, particularly OBA that involves ad networks collecting, sharing, and inferring personal information. From the open-ended responses in our study exploring the motivations for third-party use (Chapter 5), we learned that constraints in monetary and human resources often inhibit the considerations of (privacy-friendly) alternatives to popular third-party solutions that are made available free of charge but instead monetize user data.

Tracking-free subscriptions

During the manual annotation of websites for our work presented in Chapter 3, we noticed the first online editions of newspapers including Der Standard (Austria) and the Washington Post[1] (United States) to have introduced *tracking walls*, website-blocking consent dialogs that present visitors with the choice to either agree to web tracking and continue to use the page for free or purchase a subscription and get access to a (purportedly) tracking-free website. Whether such a forced decision between tracking or payment is compliant with the GDPR's requirement for "freely given" consent, is debatable [137, 196, 236]. Over the last few years, similar subscription schemes have been introduced by an increasing number of websites, presumably in reaction to the GDPR and to dwindling revenue for content creators from online advertising, with targeting ads to visitors barely making a difference [171]. These subscriptions do not necessarily involve blocking initial access to the site or services but have in common that they make claims that can be expected to reduce the amount of tracking visitors are subject to, such as "ad-free" or the explicit claim that no tracking is taking place. Irrespective of the valid argument that these business models foster a "privacy only for the rich" attitude [196], it would still be valuable to investigate if these claims hold true, as these practices currently do exist and an evaluation could inform interested individuals in their decision whether

1 At the time of writing this thesis, the Washington Post had discontinued its "tracking wall" and explicitly tracking-free subscription.

it would be worth it to pay to reduce the amount of third-party web tracking they are exposed to. Additionally, in case violations of the claims are detected, businesses could be held accountable. Whether paying for privacy is worth it has been previously explored in different contexts, including paid mobile applications [109, 246], and participants' willingness to do so was investigated in various fictitious scenarios including social networks [240]. While previous research has brushed at the privacy implications of paywalls on news websites and found no difference in tracking for a small sample of ten websites [211], it did not investigate this question at scale and did not take into account if the paid versions made any privacy-related claims. Future work could address this research gap and compare the privacy-related claims of website subscriptions to the actual tracking practices under the respective subscription plans. It has to be noted that paying for a service automatically involves transmission of sensitive personal information, including financial data, to the service, which could lead to an even higher dissemination of user data.

Browser-based micropayments

A more flexible and potentially more private option to monetize Web content could be micropayments that were first suggested in a networking context in the 1960s [31] and found renewed interest after the introduction of Bitcoin and Blockchain technology in 2009 [30]. A current suggestion for browser-based micropayments that do not require a common payment network as an intermediary is Web Monetization [289], a proposed W3C standard based on the Interledger protocol [133]. Directly funding content creators could be a first step in the direction of a less centralized Web, with less data in the hands of advertising networks and the vendors of other widely used third-party services [30].

Privacy-friendly advertising

Another option for creators of Web content to monetize their business are the non-targeted forms of advertising already mentioned in Section 2.1.5, including contextual (e. g., EthicalAds [65]) or static ads. Promoting the adoption of these alternatives to the predominant advertising networks employing OBA, which does not result in significantly higher revenue [171], could ultimately lead to a win–win situation for both websites and visitors – generating revenue while preserving the fundamental right to data privacy for everyone, not just the wealthy [196].

7.3 FUELING THE STANDARDIZATION OF WEB PRIVACY INFORMATION

Our notification study in Chapter 6 has illustrated the difficulties of reaching and communicating privacy issues to website operators. Still, as shown by previous notification campaigns about security issues, the approach per se remains promising and there is ample opportunity for future work to determine how to best address the human factor in this type of research.

Point of contact for privacy inquiries

On the technical side, web privacy research could aid regulators in enforcing existing privacy legislation by developing standards how information related to user privacy is presented on the Web. This would also facilitate future measurements on privacy mechanisms on websites, such as the prevalence of privacy policies and consent notices as we did in the work that served as the basis for Chapter 3. One idea already mentioned in Section 6.7 could be the introduction of a standardized point of information regarding a website's privacy practices,

similar to `security.txt` [86, 87]. As with many Web standards, including use of generic email addresses, large and popular websites are more likely to adopt such a proposal, unless forced to by law or developments in the browser market. Poteat and Li [218] confirmed this for `security.txt` adoption and found deployment rates between 11–16 % for the top 100 websites in the Alexa ranking, 8–10 % for the top 1,000 websites, 3-4 % for the top 10,000, and only about 1 % for the top 100,000.

Privacy policy location and format

Independent of such a file, websites could also be encouraged to provide a plaintext version of their privacy policy at a standardized location to make it easier for both human visitors and automated research efforts to find, extract, and store. This idea is now new: P3P was an early proposal that went even further and attempted to standardize how privacy information is made available on the Web in a machine-readable format, but failed due to lack of adoption. As we have seen in Chapter 6 and emphasized by Schwartz in his recapitulation why P3P failed [242], pursuing its core idea is still worthwhile, even more so under the GDPR and other recent privacy legislation that introduced new requirements for transparency and control, unwillingly increasing the cognitive burden for people whose information is processed. As pointed out for the case of automated privacy signals in the browser, regulators could make adoption of such standards mandatory – but convincing them to do so would still require widespread support by the scientific community and/or actors in the online tracking ecosystem, which can only be achieved by learning from the reasons why earlier proposals failed.

8

CONCLUSION

In this thesis, we investigated the General Data Protection Regulation's lack of a clear effect on the prevalence of third-party services on websites, as identified by the first measurement studies undertaken shortly after the law became enforceable in May 2018. We moved beyond measurements of third-party presence in an effort to better understand this lack of change, find possible other effects of the GDPR, and look into means to motivate more privacy-friendly integrations, as mandated by the GDPR's "data protection by design and by default" principle.

PREVALENCE AND IMPLEMENTATION OF CONSENT NOTICES First, we investigated in Chapter 3 if the GDPR's increased transparency requirements and the need for a legal basis for data collection led to websites providing more information about and control of their data processing practices. This was fueled by the impression that in the months before the GDPR enforcement date an increasing number of websites displayed notices asking for consent to the use of cookies. Specifically, we were guided by the following research question:

RQ 1: Has the GDPR provided website visitors with greater transparency and control regarding the collection of their personal information by third-party services on websites?

To answer it, we conducted a longitudinal measurement study of transparency mechanisms on European websites and found a significant increase in cookie consent notices before and after the GDPR enforcement date. In addition, we developed a classification of consent notices based on the options they present to website visitors. Investigating the prevalence of the different types and the capabilities of popular consent libraries, we found that only few websites provided website visitors with meaningful options to control the collection of their personal information through third-party services.

WEBSITE VISITORS' PERCEPTION OF CONSENT NOTICES Our second study in Chapter 4 directly built upon these results, in particular the observation that the majority of existing notices either did not provide any choice at all or overwhelmed website visitors with long lists of third-party vendors. We were interested in more thoroughly exploring the design space for the user interface of consent notices, including those that offer a real, non-overwhelming choice, and asked:

RQ2: How do website visitors perceive and interact with different types of consent notices, if given an actual choice to allow or deny consent?

From a sample of real-world consent notices we derived the design space for the UI of single-layer consent notices. Based on the identified parameters we designed three experiments investigating how notice position, available options

and nudging, and language and presence of a privacy policy link influenced user interaction. We conducted the first study of users' consent behavior on a real-world website and found that people prefer notices with a binary "Accept–Decline" mechanism but appreciate the availability of more fine-grained options based on categories of third-party services. Unless options are preselected, which was later ruled to violate consent requirements [69], people are unlikely to explicitly consent to data processing through third-party services. Still, we found widespread misconceptions, including websites not working properly unless consent to the use of cookies is given.

WEBSITES' CONSIDERATIONS IN THE USE OF THIRD-PARTY SERVICES In Chapter 5 we shifted our focus towards understanding the lack of change in the use of third-party services by websites. We turned to the people responsible for their integration into websites – web developers and people in similar roles – and conducted the first study that investigated the motivation behind the use of third-party services on websites, asking:

RQ3: Do people working with websites consider visitors' privacy in the integration of website functionality that is often integrated via third-party services?

In an online mixed-methods study with 395 participants we explored if and how people working with websites consider visitors' privacy in the selection and configuration of solutions to integrate a desired functionality. We found a lack of awareness of the data collection not directly associated with the core functionality of a service, including IP addresses and other technical parameters. Alternatives to the ultimately selected solutions were rarely investigated for reasons including familiarity with an existing service, and roadblocks to privacy-enhancing configurations include website creators' trust in third-party vendors and their perceived inability to have an influence on data collection. Overall, website visitors' privacy was rarely considered except for certain types of functionality for which data protection authorities had issued concrete integration guidelines.

EMAIL NOTIFICATIONS ABOUT THIRD-PARTY WEB TRACKING WITHOUT CONSENT Searching for ways to increase website owners' awareness of third parties' data collection practices, in Chapter 6 we looked into email notifications as a method previously used to raise awareness and remediation of malicious abuse of Web infrastructure and security vulnerabilities, guided by the following research question:

RQ4 Can email notifications motivate website operators to fix complex privacy issues, including use of third-party cookies without visitors' consent?

We conducted the first large-scale email notification study that informed website operators about vendor-agnostic, complex privacy issues. We notified websites about four privacy issues and one security vulnerability and compared the effect of our notifications between security and privacy issues with regard to remediation rates and recipient feedback. With one major obstacle being reachability of websites, we found only limited effect of the notification emails

on remediation rates but learned from participant feedback that privacy notifications were not as well received as those about a security vulnerability. Reasons include misconceptions of the GDPR's applicability, the data collection practices of both the website and the integrated third-party services, and difficulties in pinpointing the problem.

FUTURE WORK Finally, in Chapter 7 we encouraged the research community to find consent mechanisms that ease the cognitive burden on website visitors, help website owners with the implementation of "privacy by design and by default", and aid lawmakers and data protection authorities through proposals for the standardization how privacy information is presented on the Web and other Internet-based communication systems.

FINAL THOUGHTS Integration into websites can provide the vendors of third-party services with the opportunity to collect significant amounts of personal information about visitors, particularly if the service is used by large shares of websites, as in the case of Google services including Analytics, Ads, and Fonts or plugins for the integration of content from large social media platforms. Encouraging websites to move away from the use of widespread third-party services that rely on monetization of visitor data can be one important step towards decentralization of power and opinion through large Internet companies, currently under scrutiny in different jurisdictions including the European Union [37] and the United States [97]. This thesis is meant to contribute to this public debate and encourage further research on how to foster the use of technology that respects users' privacy while still providing them with the treasured technological amenities of the digital age.

A

APPENDICES

A.1 SURVEY: COOKIE CONSENT NOTICES

This appendix contains the survey instrument for our field study about consent notices (Chapter 4). The following tables only list the questions asked in the main questionnaire (i. e., without the intro text, consent form, and end message) and participants' responses (n = 110).

[R] indicates answers displayed in random order. All questions were non-mandatory. All questions and answer options were translated from German. E1, E2, and E3 stand for the respective experiments. Acronyms indicate the respective experimental condition:

- BIN-E1 = the binary notice shown at six different positions in Experiment 1
- NOP = no option
- CON = confirmation
- BIN = binary
- CAT = categories
- VEN = vendors
- NN = non-nudging
- NU = nudging
- TE = technical
- NT = non-technical
- PP = privacy policy link
- NP = no privacy policy link

Table A.1: Motivation for (Not) Interacting With the Cookie Consent Notice

.

Q1-clicked[a]: You just clicked the cookie consent notice[b] on the website [WEBSITE_NAME].
Which of the following statements describe your motivation to click the notice?
I clicked the cookie consent notice[b] ... [multiple choice]

	E1	E2	E3	Total	%
... to protect me from dangers from the Internet.[R]	0	3	3	6	9.8
... to protect my privacy on the Internet.[R]	0	5	6	11	18.0
... because the website does not work otherwise.[R]	2	11	3	16	26.2
... to see fewer ads.[R]	1	1	3	5	8.2
... out of habit.[R]	1	10	2	13	21.3
... because the notice distracts me from viewing the website.[R]	6	25	13	44	72.1
Other: [free text]	0	0	1	1	1.6
I do not know why I clicked the notice.	1	1	1	3	4.9
I prefer not to answer.	0	0	0	0	0
# Answers	11	56	32	99	
# Participants	8	34	19	61	

Q1-notclicked[a]: You did not click the cookie consent notice[b] on the website [WEBSITE_NAME].
Which of the following statements describe your motivation to not click the notice?
I did not click the cookie consent notice ... [multiple choice]

	E1	E2	E3	Total	%
... because I have not noticed it.[R]	4	11	5	20	40.8
... because it did not offer enough choices.[R]	0	0	3	3	6.1
... because I do not know what happens if I click the notice.[R]	1	6	4	11	22.4
... because I think that my selection does not have any effect.[R]	1	4	4	9	18.4
... because I do not know what cookies are.[R]	0	2	0	2	4.1
... because I do not care which cookies the website uses.[R,c]	1	3	2	6	12.2
... Other: [free text]	1	10	2	13	26.5
... I do not know why I did not click the cookie consent notice.	1	0	0	1	2.0
... I prefer not to answer.	0	2	0	2	4.1
# Answers	9	38	20	67	
# Participants	8	26	15	49	

[a]Q1-clicked and Q1-notclicked were only displayed to participants who had clicked / had not clicked the notice, respectively.
[b]In Experiment 3, "cookie consent notice" was changed to "privacy notice" in the conditions *Non-Technical–PP Link* (NT-PP) and *Non-Technical–No PP Link* (NT-NP).
[c]In Experiment 3, this answer was changed to "because I do not know what data this is about" in the conditions *Non-Technical–PP Link* (NT-PP) and *Non-Technical–No PP Link* (NT-NP).

Table A.2: Expectation of the Website's Data Collection

.

Q2: What do you think – what data does the website [WEBSITE_NAME] collect about you when you access the website? [multiple choice]	E1	E2	E3	Total	%
The posts I am reading on the website.[R]	10	40	17	67	60.9
My residence.[R]	6	14	7	27	24.5
The links I click on the website.[R]	14	45	27	86	78.2
My IP address.[R]	11	39	22	72	65.5
The device I am using to access the website.[R]	10	36	19	65	59.1
The website does not collect any data about its visitors.[R]	0	4	1	5	4.5
My name.[R]	2	9	3	14	12.7
Other websites I visit besides [WEBSITE_NAME].[R]	5	17	10	32	29.1
Other: [free text]	3	2	1	6	5.5
I prefer not to answer.	0	0	0	0	0
# Answers	61	206	107	374	
# Participants	16	60	34	110	

Table A.3: Perception of the Cookie Consent Notice Displayed to the Participant

.

This is the cookie consent notice[b] the website has shown you. [IMAGE]
Please rate the following statements about this notice.

Q3: I think the number of choices offered by the above cookie consent notice[b] is ...

	Exp. 1	Exp. 2									Exp. 3			
	BIN-E1	NOP	CON-NN	CON-NU	BIN-NN	BIN-NU	CAT-NN	CAT-NU	VEN-NN	VEN-NU	TE-PP	TE-NP	NT-PP	NT-NP
... too low	9	3	3	5	3	1	1	2	1	2	1	1	0	1
... just right	7	1	0	3	7	3	2	3	1	2	4	8	6	6
... too high	0	0	1	1	0	0	3	2	0	3	2	0	3	0
... No answer	0	2	0	1	2	0	1	0	1	0	0	0	2	0
Total	16	6	4	10	12	4	7	7	3	7	7	9	11	7

Q4: The above cookie consent notice[b] allows me to control the website's behavior.

	Exp. 1	Exp. 2									Exp. 3			
	BIN-E1	NOP	CON-NN	CON-NU	BIN-NN	BIN-NU	CAT-NN	CAT-NU	VEN-NN	VEN-NU	TE-PP	TE-NP	NT-PP	NT-NP
Strongly disagree	6	3	3	0	2	0	1	1	0	1	1	1	0	0
Somewhat disagree	3	2	0	3	3	2	2	1	1	3	2	0	3	0
Neutral	6	0	1	4	3	0	0	1	1	1	1	1	4	2
Somewhat agree	1	1	0	2	4	1	4	3	1	1	1	4	4	5
Strongly agree	0	0	0	0	0	1	0	1	0	1	2	2	0	0
No answer	0	0	0	1	0	0	0	0	0	0	0	1	0	0
Total	16	6	4	10	12	4	7	7	3	7	7	9	11	7

Q5[d]: I think the decision which option to select in the cookie consent notice[b] is ...

	Exp. 2				Exp. 3			
	CAT-NN	CAT-NU	VEN-NN	VEN-NU	TE-PP	TE-NP	NT-PP	NT-NP
... very easy	2	0	1	1	4	3	4	1
... easy	0	2	1	0	0	2	2	2
... neither easy nor hard	2	2	0	2	2	2	5	4
... hard	2	2	1	2	1	1	0	0
... very hard	1	1	0	2	0	0	0	0
No answer	0	0	0	0	0	1	0	0
Total	7	7	3	7	7	9	11	7

[b] In Experiment 3, "cookie consent notice" was changed to "privacy notice" in the conditions *Non-Technical–PP Link* (NT-PP) and *Non-Technical–No PP Link* (NT-NP).
[d] Q5 was only shown to participants who had seen a category- oder vendor-based notice on the website.

Table A.4: Perception of the Cookie Consent Notice Displayed to the Participant (cont.)

Q6[d]: Please explain your answer to the previous question. [free text]					
Code	*Explanation*	E2	E3	Total	%
Transparent	The participant considers the consent notice to be transparent.	1	5	6	15.8
Privacy	The participant's preferences are privacy-focused, i. e., the least invasive option is chosen.	2	5	7	18.4
Options clear	The options offered by the consent notice are considered clear / easy to understand.	0	3	3	7.9
Options unclear	The options offered by the consent notice are considered unclear / not easy to understand.	4	2	6	15.8
Notice clear	The participant expressed that the mechanism was clear, but did not specify which part.	1	3	4	10.5
Notice unclear	The participant expressed that the mechanism was unclear but did not specify which part.	2	0	2	5.3
Too complicated	The consent notice was considered too complex.	4	1	5	13.2
Don't care	The participant stated they did not care which cookies the website used.	3	0	3	7.9
Other		4	2	6	15.8
# Participants		60	34	94	

[d]Q6 was only shown to participants who had seen a category- oder vendor-based notice on the website.

Table A.5: General Understanding of Cookie Consent Notices

.

This is another cookie consent notice. [Image of the binary notice in Figure 4.1 (a) (bb)]						
Q7: What do you think happens when you click "Decline"? [free text]						
Code	Explanation	E1	E2	E3	Total	%
Site blocked	The content of the website cannot be accessed at all.	6	13	9	28	29.8
Functionality limited	The content of the website can be viewed, but some parts may not work.	2	10	5	17	18.1
Site accessible	The content of the website can be accessed.	0	3	1	4	4.3
No data collected	The website visitor's personal data is not collected or processed.	2	4	5	11	11.7
No cookies set	The website does not store any cookies in the visitor's browser.	1	8	3	12	12.8
Fewer ads	The website displays fewer or no ads.	0	3	2	5	5.3
Notice	The participants only mentions effects regarding the consent notice.	0	2	3	5	5.3
No change	Declining cookies does not have any effect.	4	7	1	12	12.8
Don't know		2	0	1	3	3.2
Other		0	2	4	6	6.4
# Participants		15	51	28	94	
Q8: What do you think happens when you click "Accept"? [free text]						
Code	Explanation	E1	E2	E3	Total	%
Data collected	The participant's personal data is collected and / or processed.	9	10	10	29	30.9
Cookies stored	Cookies are stored in the user's browser.	4	9	6	19	20.1
Site accessible	The content of the website can be accessed.	0	16	5	21	22.3
Notice	The participants only mentions effects regarding the consent notice.	0	3	2	5	5.3
Ads	The participant is subject to advertising.	6	11	6	23	24.5
Profiling	The participant's personal data is used to create a profile of their interests.	5	8	6	19	20.2
Other purposes	The participant's personal data is used for other purposes.	2	0	2	4	4.3
No change	Clicking "Accept" does not have any effect.	0	4	3	7	7.4
Don't know		0	3	0	3	3.2
Other		0	3	1	4	4.3
# Participants		15	51	28	94	

A.2 SURVEY: WEB TECHNOLOGIES – SELECTION, INTEGRATION, AND CONFIGURATION

This appendix contains the survey instrument for the study about considerations in the use of third-party services presented in Chapter 5. Presented below is only the main questionnaire as shown in Figure 5.1, i. e., without the intro text, privacy policy, debriefing, and end message. Except for Q2-0 in the *GitHub–Mandatory* condition, all questions were non-mandatory.

Survey Title

Web Technologies: Selection, Integration, and Configuration

1. Your Background

First we would like to learn about your background and your work on websites. Throughout this survey, by "work on websites" we mean your involvement to some degree in the design, development, deployment, maintenance, and/or management of a website.

1-1 How many websites have you worked on in the last 3 years? [single choice]

- 0
- 1
- 2–5
- 6–10
- 11–25
- 26–50
- 51–100
- $>$ 100

1-2 What is your current employment status with regard to your work on websites? [multiple choice]

- Full-time employment
- Part-time employment
- Self-employed / freelancer
- Intern
- Hobbyist
- Unemployed
- Retired
- Unable to work
- Other: [free text]
- Prefer not to say

1-3 Below is a list of functionalities often found on websites. Which of these functionalities have you previously worked with on websites? [multiple choice; order of answers randomized]

- Advertising (e. g., banner ads, video ads, content recommendation, affiliate links)
- Customer / user interaction (e. g., user comments, contact forms, chat, mailing lists)
- Embedded media (e. g., video, audio, maps, slideshows)
- Front-end libraries or design resources (e. g., non-standard fonts, CSS frameworks, JavaScript libraries)
- User login / authentication
- Payment systems
- Privacy popups / privacy forms (e. g., cookie consent notices, CCPA "Do Not Sell")
- Website protection (e. g., anti-spam, bot mitigation techniques)
- Social media integration (e.g., social media buttons, widgets, embedded feeds)
- Web analytics (e. g., page visits, heatmaps, session replay)

2a. Website

To learn more about your experience with different web technologies, the rest of the survey will ask you about a specific website you have recently worked on.

2-0 Please name one website you recently worked on, i. e., you were involved in the design, development, deployment, maintenance, or management of that website, and that you remember well.

(*If recruited via website:* Ideally, this is the website through which we contacted you, which is mentioned in the email invitation to this survey. If you were not in any way involved in the design, development, deployment, maintenance, or management of that website, you are welcome to provide another website you recently worked on.)

We will keep this website – and any other information that could identify you – confidential and only share it with involved researchers.

Please enter the website's web address below, including the top-level domain (e. g., `youtube.com`, `guardian.co.uk`).

For the remainder of this survey, all questions are going to refer to this website as "the website." [free text]

(In the GitHub–Mandatory condition, we required participants to enter something but did not check if it was a valid URL.)

2b. Website Info

In Part 2 of the survey, we would like to learn some more information about the website you just named.

2-1 What is / are your role(s) with regard to the website? [multiple choice]

- Product or project manager
- Content creator or contributor
- Social media manager
- Marketing
- Sales
- Quality assurance
- User experience
- (Web) developer, programmer, or software engineer
- Administrator or (web) operator
- Legal counsel
- Data protection officer
- Customer service / customer support / customer relations
- Other: [free text]

2-2 What is roughly the size of the team working on the website, i. e., how many people have been involved in the website's design, development, deployment, maintenance, and management? [single choice]

- I am the only team member
- 2–5
- 6–10
- 11–25
- 26–50
- 51–100
- $>$ 100
- Don't know

2-3 Please select which country the company or organization operating the website is based in. If the company or organization has sites in multiple countries, please select the country in which the company or organization's headquarters are located. [single choice, answer options: dropdown list with names of all countries]

2-4 What regions or countries is the website targeting or being used in? [free text]

2-5 What is the website's revenue model? [multiple choice]

- Targeted advertising (e. g., ad networks)
- Non-targeted advertising (e. g., contextual or static ads)
- Affiliate marketing / affiliate links
- Donations
- Subscriptions / membership
- Sponsored posts / articles
- Products / services sold on the website
- Supported by other revenue streams (i.e., goods or services not directly sold on the website)
- Other: [free text]
- Not applicable (website does not have a revenue model)
- Don't know

2-6 Which of the following features or functionalities are used on the website? [single choice for each, answer options: Yes / No / Not sure]

- Advertising (e. g., banner ads, video ads, content recommendation, affiliate links)
- Customer / user interaction (e. g., user comments, contact forms, chat, mailing lists)
- Embedded media (e. g., video, audio, maps, slideshows)
- Front-end libraries or design resources (e. g., non-standard fonts, CSS frameworks, JavaScript libraries)
- User login / authentication
- Payment systems
- Privacy popups / privacy forms (e. g., cookie consent notices, CCPA "Do Not Sell")
- Website protection (e. g., anti-spam, bot mitigation techniques)
- Social media integration (e. g., social media buttons, widgets, embedded feeds)
- Web analytics (e. g., page visits, heatmaps, session replay)

2-7 For each of the following functionalities present on the website, how involved have you been regarding their integration into the website? [list of all functionalities tagged with "Yes" in previous question, single choice for each, answer options:]

- I decided how to integrate this functionality
- I integrated / implemented this functionality
- I maintain or manage the integration of this functionality
- I have not been involved in the integration of this functionality

3. Integration of Website Functionalities (category-specific)

In Part 3 we would like to ask you a few questions about the integration of some of the functionalities you indicated to have worked with on the website. You will be shown these questions for at most three different functionalities, regardless of how many you have selected in the previous question.
(For up to three categories randomly selected from those the participant has indicated involvement in the previous question, they are asked the following questions.)
You indicated that you have been involved to some degree in the integration of [FUNCTIONALITY (examples)] on the website. Now we would like to ask you a few more questions about how this functionality has been integrated.

3-1 For which purposes or use cases is [FUNCTIONALITY] technology used on the website? [free text]

3-2 a. *(Generic:)* Which technology has been used to integrate [FUNCTIONALITY] into the website? If the website uses multiple technologies for this, please consider all of them combined (your "solution") when answering the following questions. [multiple choice + free text]

- We developed it ourselves
- We installed a third-party software on the website's host system (please name software:) [free text]
- We integrated an external third-party service (please name service:) [free text]
- Other (please specify:) [free text]
- Don't know

b. *(Payment:)* What kind of payment service(s) does the website use? [multiple choice + free text]

- Payment method(s) that do not require other parties for processing (e. g., cash, gift cards) (please name method(s):) [free text]
- Service(s) that only involve banks on either side (e. g., bank transfer, Lastschrift) (please name service(s):) [free text]
- Service(s) that involve third parties (e. g., credit card, PayPal) (please name service(s):) [free text]
- Other (please specify:) [free text]
- Don't know

c. *(Embedded Media:)*

i. What type of embedded media does the website use? [multiple choice]

* Embedded maps
* Embedded videos

- Embedded audio
- Other (please specify:) [free text]
- Don't know

ii. (1) *(If map, audio, or video:)* You indicated that the website uses embedded (maps | videos | audio).

(a) Where are these (map | video | audio) resources hosted? [multiple choice]

- The (map | video | audio) resources are hosted on the website's host system
- The (map | video | audio) resources are hosted with a third-party service (please name service:) [free text]
- Other (please specify:) [free text]
- Don't know

(b) *(If map, audio, or video and third-party hosting:)* How are these externally hosted (map | video | audio resources) embedded into the website? If the website uses multiple technologies for this, please consider all of them combined (your "solution") when answering the following questions. [multiple choice]

- Embedding code provided by the third party that hosts the resources
- Embedding code provided by another third-party service (please specify service:) [free text]
- Embedding code we have written ourselves
- Other (please specify:) [free text]
- Don't know

(2) *(If "Other":)* You indicated that the website uses some other kind of embedded content. How is this content integrated into the website? If the website uses multiple technologies for this, please consider all of them combined (your "solution") when answering the following questions. [free text]

d. *(Social Media:)*

i. What type of social media integration does the website use? [multiple choice]

- Profile buttons or links
- Share buttons or widgets
- Embedded posts or feeds
- Other: [free text]
- Don't know

ii. (1) *(If profile / share buttons or embedded:)* You indicated that the website uses (buttons or links to social media profiles | social media share buttons or widgets | embedded social media posts or feeds). Which technology has been used to integrate them into the website? If the website uses multiple technologies for this, please consider all of them combined (your "solution") when answering the following questions. [multiple choice]

- Code we have written ourselves
- Code provided by social media site(s)
- Code or plugin provided by another third-party service (please specify service:) [free text]
- Other (please specify:) [free text]
- Don't know

(2) *(If "Other":)* You indicated that the website uses some other kind of social media integration. How is it integrated into the website? If the website uses multiple technologies for this, please consider all of them combined (your "solution") when answering the following questions. [free text]

3-3 *(If involved in selection:)* You indicated that you were involved in deciding how [FUNCTIONALITY] was integrated into the website. Please describe why this specific type of integration or this particular service was selected. [free text]

3-4 *(If involved in selection:)*

a. *(Generic:)* When making this decision, were other ways for integrating [FUNCTIONALITY] into the website considered? [multiple choice]

- We considered a solution we have developed (or were going to develop) ourselves
- We considered (another) third-party software installed on the website's host system (please name software:) [free text]
- We considered a(nother) service hosted with a third party (please name service(s):) [free text]
- We directly decided to use the current solution
- Other (please specify:) [free text]
- Don't know

b. *(Payment:)* When making this decision, were other ways for integrating payment systems into the website considered? [multiple choice]

- We considered (other) methods that do not include any other party (e. g., cash, gift cards) (please name method(s):) [free text]

- We considered service(s) that only involve banks on either side (please name service(s):) [free text]
- We considered (other) service(s) that involve third parties (please name service(s):) [free text]
- We directly decided to use the current solution
- Other (please specify:) [free text]
- Don't know

c. *(Embedded Media:)* When making this decision were other ways for integrating embedded media into the website considered? [multiple choice]
   - We considered self-hosting the embedded media resources
   - We considered hosting the embedded media resources with a(nother) third party (please specify service:) [free text]
   - We considered embedding code provided by the third-party service that hosts the resources (please specify service:) [free text]
   - We considered embedding code provided by a different third-party service (please specify service:) [free text]
   - We considered embedding code we have written (or were going to write) ourselves
   - We directly decided to use the current solution
   - Other (please specify:) [free text]
   - Don't know

d. *(Social Media:)* When making this decision, were other ways for integrating social media into the website considered? [multiple choice]
   - We considered a solution we have developed (or were going to develop) ourselves
   - We considered code provided by the social media site(s)
   - We considered a solution provided by a different third-party service (please specify service:) [free text]
   - We directly decided to use the current solution
   - Other (please specify:) [free text]
   - Don't know

3-5 *(If involved in selection:)* Why were other ways to integrate [FUNCTIONALITY] into the website (not) considered? [free text]

3-6 *(If involved in selection:)* Which sources of information did you use to select a solution to integrate [FUNCTIONALITY] into the website? [multiple choice]

- The website's team
- Professional network (people external to the website team)
- Private network (e. g., friends)
- Sales representative of third-party software / service
- Official website(s) / documentation of third-party software / service
- Legal documents by third-party software / service (e. g., terms of service, privacy policy)
- Online blogs / magazine articles
- Online discussion forums (e. g., Reddit, StackOverflow)
- Other: [free text]

3-7 *(If involved in implementation or maintenance:)* Which sources of information did you use to configure the [FUNCTIONALITY] solution on the website? [multiple choice, same answer options as in Q3-6]

3-8 *(If not involved in selection:)* You indicated that you were not involved in the decision how to integrate [FUNCTIONALITY] into the website. Who decided how [FUNCTIONALITY] should be integrated into the website? [multiple choice]

- Product or project manager(s)
- Content creator(s) or contributor(s)
- Social media manager(s)
- Marketing
- Sales
- Quality assurance
- User experience
- (Web) developer(s), programmer(s), or software engineer(s)
- Administrator(s) or (web) operator(s)
- Legal counsel(s)
- Data protection officer(s)
- Customer service / customer support / customer relations
- CEO and/or other upper level management
- Investor(s)
- Other: [free text]
- Don't know

3-9 Overall, how satisfied are you with the [FUNCTIONALITY] integration solution on the website, with regard to the following criteria? [single choice for each of the following, answer options: Very satisfied, Satisfied, Neither satisfied nor dissatisfied, Dissatisfied, Very dissatisfied, Don't know]

- Visitors' privacy
- Ease of integration
- Ease of use for visitors
- Performance (e. g., page speed)
- Features meet requirements

4. Data Practices of Website Functionalities (category-specific)

In Part 4 of the survey, we would like to learn more about your experience with the data practices of the technologies we just asked you about in Part 3.
(The following questions are asked for each functionality for which the participant has also seen Part 3.)

4-1 *(If third-party service is used to implement [FUNCTIONALITY]:)* Sometimes third-party services, when integrated into a website, collect information about the website's visitors, either to provide the service or for their own / other purposes. To the best of your knowledge, what information about the website's visitors does the third-party solution used for [FUNCTIONALITY] collect?

[Items taken from the "Information Type" section of the annotation scheme for the OPP-115 corpus of privacy policies [295]; single choice for each, answer options: Yes, No, Unsure]

- Financial information (e. g., credit or debit card data, credit scores)
- Health, genetic, or biometric data
- Contact information (e. g., name, email address, phone number)
- Location (e. g., GPS location, postal code)
- Demographic data (e. g., gender, age, education)
- Personal identifiers (e. g., social security, ID card or driver's license number)
- User online activities (e. g., pages visited, time spent on pages)
- User profile on the website (e. g., profile settings, data the user has uploaded to the website)
- Social media data
- IP address or device IDs
- Cookies or other tracking elements
- Device information (e. g., browser or operating system used by website visitors)

4-2 *(If involved in implementation or maintenance:)* Did you make any specific effort(s) to protect the website's visitors' privacy when configuring the [FUNCTIONALITY] solution on the website? [single choice]

- Yes
- No
- Don't know

4-3 a. *(If yes:)* Please describe which efforts you have made and why. [free text]

b. *(If no:)* Please describe why you did not make any specific efforts. [free text]

5. Demographics

Finally, we would like to ask you some basic demographic questions to better understand who participated in our study.

5-1 What is your age (in years)? [single choice]

- 18–24
- 25–34
- 35–44
- 45–54
- 55–64
- 65–74
- 75+
- Prefer not to disclose

5-2 What is your gender?[1] [multiple choice]

- Woman
- Man
- Nonbinary
- Prefer to self-describe: [free text]
- Prefer not to disclose

5-3 What is the highest educational degree you have completed? [single choice]

- No schooling completed
- Some high school, no diploma
- High school graduate, diploma, or equivalent (e. g., GED, Abitur, baccalauréat)
- Some college credit, no degree

1 As recommended by Spiel et al. [253].

- Trade / technical / vocational training
- Associate degree
- Bachelor's degree
- Master's degree or equivalent (e. g., German Diplom)
- Professional degree (e. g., JD, MD, German Staatsexamen)
- Doctoral degree (e. g., PhD)
- Other: [free text]
- Prefer not to disclose

5-4 In what field(s) did you receive your degree or vocational training?[2] [multiple choice]

- Computer and information sciences
- Mathematics
- Engineering
- Life sciences (e. g., biology, health sciences, medicine)
- Social sciences / social work / human services
- Education
- Law
- Psychology / behavioral science
- Business / economics
- Liberal arts / humanities
- Art / music
- Journalism
- Vocational
- Other: [free text]
- Not applicable
- Prefer not to disclose

5-5 Have you ever received any kind of training or educated yourself on data protection or privacy? [single choice]

- Yes (please specify:) [free text]
- No
- Prefer not to disclose

2 Adapted from a Pew Research survey [214], using the subcategories for some fields.

A.3 NOTIFICATION EMAILS

This appendix contains the text of the notification emails we sent in our notification study in Chapter 6.

Sender / Display Name

Matthias Michels | CISPA Helmholtz Center for Information Security

Subject

[Security and] data protection issue[s] on your website [DOMAIN] *or* Security issue on your website [DOMAIN]

Email Text

Hello,

We are a group of security and privacy researchers from the CISPA Helmholtz Center for Information Security and Ruhr University Bochum in Germany. As part of our current research project, we analysed potential security and data protection issues on websites.

We would like to raise your attention to the following security and data protection issue(s) on your website [DOMAIN]. Please note that we do not offer a conclusive legal assessment or consultancy on an individual website's legal compliance.

NO PRIVACY POLICY. For public websites that use European domains, are hosted in the EU, or may be used by European users, any collection of users' personal data is governed by the EU General Data Protection Regulation (GDPR). If a website meets these conditions, the operator is legally required by Article 13 of the GDPR to have a privacy policy explaining the use of their visitors' personal data. Personal data also encompasses the processing of communications data such as IP addresses of users even if no additional information is collected. The privacy policy has to inform users about the use of their personal data in a concise, transparent, intelligible, and easily accessible form.

Our automated analysis of your website did not detect a privacy policy, which may indicate noncompliance with the GDPR's information requirements.

INPUT FIELDS FOR PERSONAL INFORMATION WITHOUT HTTPS. Article 32 of the GDPR requires data controllers such as website owners to implement appropriate technical and organisational measures to ensure a level of security appropriate to the risk, taking into account the state of the art. Protection of users‘ communication and interactions with your website via HTTPS is considered state of the art in data security.

Our automated analysis detected input fields on your website that allow users to enter personal data without using HTTPS secure communication to prevent eavesdropping and phishing. This may indicate noncompliance with the GDPR's data security requirements.

USE OF THIRD-PARTY COOKIES WITHOUT CONSENT NOTICE. Under Article 5 Paragraph 3 of the EU ePrivacy Directive (Directive 2009/136/EC) and respective implementations of the Directive into national law of the EU member states, the setting of individual cookies on the user's terminal equipment that are not strictly necessary for the functioning of the website is only allowed if the user has given his or her prior consent.

Our automated analysis did not detect such a consent form for the third-party cookies on your website. This may indicate noncompliance with EU ePrivacy requirements.

THIRD-PARTY COOKIES SET BEFORE INTERACTION WITH CONSENT NOTICE. Under Article 5 Paragraph 3 of the EU ePrivacy Directive (Directive 2009/136/EC) and respective implementations of the Directive into national law of the EU member states, the setting of individual cookies on the user's terminal equipment that are not strictly necessary for the functioning of the website is only allowed if the user has given his or her prior consent. Such consent has to be given in advance via a meaningful interaction by the user.

According to our automated analysis, your website does provide users with a cookie notice or consent form, but the cookies are set before any meaningful interaction of a user with the consent form takes place. This lack of explicit consent may indicate noncompliance with EU ePrivacy requirements.

PUBLICLY ACCESSIBLE GIT REPOSITORY. If the configuration folder for Git (`.git`) is reachable through HTTP, an attacker may copy the content of this repository. This allows an attacker to access the source code versioned in this repository, including any credentials or other sensitive data possibly stored there. Our automated analysis detected a publicly accessible Git repository on your website. Note that we only check for the existence of a repository and do not attempt to download any actual content. Hence, we cannot state if it contains any sensitive information.

If in Warning condition:

- *If Git:* Please note: In the worst case, access to configuration files with credentials could lead to an attacker taking over your entire website.
- *If No Privacy Policy or No HTTPS:* Noncompliance with GDPR requirements could lead to fines of up to 10 million euros or up to 2 percent of the global turnover of the preceding fiscal year according to Article 83 Paragraph 4 GDPR.[3]
- *If No Consent or Before Consent:* Fines for noncompliance with ePrivacy requirements may vary depending on national laws.

You can review more detailed information about the security and data protection issues and their remediation status on your website by visiting our web

3 Due to oversight we did not differentiate in the warning text between the two tiers of fines in Article 83 GDPR: While not having a privacy policy (violates Article 13) is subject to the fines in Article 83(5) (20 million euros / 4 % of annual turnover), non-use of HTTPS (violates Article 32) falls under Article 83(4) (10 million euros / 2 % of turnover). We do not expect this difference in maximum fines to have any significant impact on notification recipients' remediation behavior.

interface at `https://notify.cispa.de/reports/[DOMAIN]/report-[UNIQUE_ID]`.

Since this notification is part of an ongoing research project, we will recheck your website to verify if the issues have been fixed. If you wish us to stop this check, please visit our web interface at `https://notify.cispa.de/reports/[DOMAIN]/report-[UNIQUE_ID]` to opt out or contact us at `info@notify.cispa.de`.

Help us improve our notification process with anonymous feedback at: `https://notify.cispa.de/reports/[DOMAIN]/report-[UNIQUE_ID]/notification-survey`.[4]

Should you need further information or have any other questions, please do not hesitate to contact us using the same email address.

Best regards,
Matthias Michels
Security Researcher
CISPA Helmholtz Center for Information Security
Stuhlsatzenhaus 5
66123 Saarbrücken
Germany

A.4 NOTIFICATION STUDY INFO WEBSITE

This appendix contains the text of the info website that informed notification recipients about the research project behind the study presented in Chapter 6 and the involved institutions.

Website Text

We are security and privacy researchers from the Secure Web Applications Group (`https://swag.cispa.saarland/`) at the CISPA Helmholtz Center for Information Security and the Systems Security group (`https://informatik.rub.de/syssec/`) at Ruhr-Universität Bochum, both in Germany. We are currently conducting a research project on large-scale security and data protection notifications. With our notifications we would like to help website owners identify and fix security and data protection issues on their websites.

Our analysis tool checks websites for the presence of a privacy policy and a cookie consent notice, whether third-party cookies are being set before consent, potentially unprotected personal information in input fields, and publicly accessible code versioning repositories. If our tool detects an issue, we notify the website owner about it via e-mail. The checks are performed in a non-intrusive

4 Clicking this link triggered a redirect to the survey. The `UNIQUE_ID` was only used to look up the notification issues, study conditions, and email group associated with the website, which were then translated into URL parameters for the survey link. No unique identifier was passed to the survey.

way. Our tool will never try to exploit a vulnerability on your server or interfere with your services.

In case you would like to contact us about this research, you can send an email to `info@notify.cispa.de`. If you want your websites to be excluded from our analysis, you can email us the domains, IP addresses, or IP ranges which should be excluded. Alternatively, if you have received an individual report for your website from us, you can use the opt-out buttons in that report.

A.5 SURVEY: SECURITY AND DATA PROTECTION NOTIFICATIONS

This appendix presents the questionnaire for the survey we invited every notification recipient to take in our notification study in Chapter 6, along with response counts.

Survey Title

Survey on Security and Data Protection Notifications

Intro Text

We are security and privacy researchers from the CISPA Helmholtz Center for Information Security and Ruhr University Bochum in Germany. In our current research we are trying to better understand how to notify websites about security and data protection issues. We recently emailed you a security and data protection notification from notify@notify.cispa.de.

You can help us improve our notification process through completing this survey. The survey is short and anonymous, and all questions are optional, so please answer the ones that you feel comfortable with. Your feedback is very valuable to us and we really appreciate your time.

Privacy Policy & Consent

We take great care in protecting our survey participants' privacy in accordance with the provisions of the General Data Protection Regulation (GDPR). Your answers to this survey will be stored securely on a server hosted by Ruhr University Bochum, Germany. Any of the survey data will only be accessible by the researchers involved in this project and will not be correlated with other data or otherwise used to identify individual participants. If we make data from this research available to the research community or the interested public, we will only publish it in an aggregated form that does not allow anyone to identify you or the website for which we sent you a notification email. You can find the contact information of the responsible data protection officers at `https://notify.cispa.de/privacy_en.html`.

Participation

Your participation in this research is completely voluntary. Once you have started the survey, you may cancel at any time by clicking the "Exit and clear

survey" button in the upper right part of the screen, and your answers will be discarded.

Contact Information

If you have any questions, comments, or concerns about the study either before, during, or after participation, please contact us at `info@notify.cispa.de`.

Questionnaire

First we would like to ask you about the security and data protection issues we found on your website. Here is a list of the issues we found: *[of the following, only the detected issues were shown]*

- No privacy policy
- Use of third-party cookies without consent notice
- Third-party cookies set before interaction with consent notice
- Input fields for personal information without HTTPS
- Publicly accessible Git repository

Table A.6: Notification questionnaire and responses.

Q1: Do you think our report is correct regarding each of the detected issues? [list of detected issues as shown above; single choice for each]

	No Privacy Policy		No Consent		Before Consent		No HTTPS		Git	
	n	%	n	%	n	%	n	%	n	%
Yes	21	70.0	10	45.5	17	63.0	1	25.0	123	87.9
No	5	16.7	5	22.7	2	7.4	1	25.0	8	5.7
Uncertain	2	6.7	4	18.2	8	29.6	2	50.0	8	5.7
N/A[a]	2	6.7	3	13.6	0	0.0	0	0.0	1	0.7
# Displayed[b]	30		22		27		4		140	

Q2: Were you aware of this / these issue(s) before we contacted you? [single choice]

	Security		Privacy		All	
	n	%	n	%	n	%
Yes	18	12.9	21	28.4	39	18.4
No	114	81.4	42	56.8	154	72.6
Don't know	7	5.0	6	8.1	13	6.1
N/A	1	0.7	5	6.8	6	2.8
# Displayed	140		74		212	

[a] Throughout the questionnaire, "Displayed" indicates how many participants had seen each question.
[b] "N/A" indicates how many of them did not provide an answer.

Table A.7: Notification questionnaire and responses (cont.).

Q3: Are you planning to make any changes to the website after receiving our message? [single choice]

	Security		Privacy		All	
	n	%	n	%	n	%
Yes	127	90.7	48	64.9	173	81.6
No	4	2.9	17	23.0	21	9.9
Don't know	7	5.0	3	4.1	10	4.7
N/A	2	1.4	6	8.1	8	3.8
# Displayed	140		74		212	

Q4 *(If "No")*: Why are you not planning to make any changes? [free text, multiple codes per answer possible]

	n	%
Non-applicability of privacy law (generic)	1	4.8
Non-EU	4	19.0
No third-party cookies used	3	14.3
No personal data collected	7	33.3
Other	5	23.8
N/A	1	4.8
# Displayed	21	

Q5 *(If any privacy issue was detected)*: Do you think the European Union's General Data Protection Regulation (GDPR) applies to your website? [single choice]

	n	%
Yes	47	63.5
No	15	20.3
Don't know	7	9.5
N/A	5	6.8
# Displayed	74	

Q6 *(If "No")*: Why do you think the GDPR does not apply to your website? [free text]

	n	%
Non-EU	6	40.0
No personal data collected	7	46.7
Other	1	6.7
N/A	1	6.7
# Displayed	15	

Table A.8: Notification questionnaire and responses (cont.).

Q7 *(If any privacy issue was detected)*: In the past, did you already make changes to the website because of the GDPR or other privacy legislation? [single choice]

	n	%
Yes	32	43.2
No	36	48.6
Don't know	1	1.4
N/A	5	6.8
# Displayed	74	

Q8 *(If "Yes")*: What changes did you make because of this privacy legislation? [free text; multiple codes per answer possible]

	n	%
Made changes to privacy policy	5	15.6
Installed cookie plugin or banner	13	40.6
Removed third-party service/cookies	6	18.8
Enforced HTTPS	2	6.3
Other	9	28.1
N/A	6	18.8
# Displayed	32	

Q9 *(If any privacy issue was detected)*: Were you aware of potential fines mandated by the GDPR before you received our message? [single choice]

	n	%
Yes, since you emailed me	9	12.2
Yes, even before you emailed me	38	51.4
No, I'm not aware of them	20	27.0
N/A	7	9.5
# Displayed	74	

Q10 *(If either "Yes" option was selected)*: In which way did this knowledge of fines influence your decision to fix the issue(s)? [free text; single code per answer]

	n	%
Reported influence of fines	13	27.7
No reported influence of fines	13	27.7
Unrelated answer	6	12.8
N/A	15	31.9
# Displayed	47	

Table A.9: Notification questionnaire and responses (cont.).

Q11: What type of support would you find helpful to fix the issue(s) we found on your website? [free text; multiple codes per answer possible]	n	%
Info in notification was sufficient	80	37.7
Better documentation of checks	35	16.5
More information about fixes	26	12.3
Other	6	2.8
N/A	83	39.2
# Displayed	212	

Q12: What is / are your role(s) with regard to the website we notified you about? [multiple choice; answers shown in random order except for "Other"]	n	%
Product or project manager	29	13.7
Content creator or contributor	42	19.8
Social media manager	8	3.8
Marketing	11	5.2
Sales	7	3.3
Quality assurance	12	5.7
User experience	11	5.2
(Web) developer, programmer, or software engineer	121	57.1
Administrator or (web) operator	128	60.4
Legal counsel	6	2.8
Data protection officer	21	9.9
Customer service / customer support / c. relations	11	5.2
Other: [free text]	21	9.9
N/A	11	5.2
# Displayed	212	

Q13: Is there anything you want to tell us about our checks, notifications, or any other issue related to this research or to security and data protection notifications in general? [free text; multiple codes per answer possible]	n	%
Positive sentiment / thanks	51	24.1
Negative sentiment	4	1.9
Email first seemed suspicious	9	4.2
More information required	5	2.4
Translation suggested	3	1.4
Tool for self-checks desired	4	1.9
Other	7	3.3
N/A	136	64.2
# Displayed	212	

End Message

Thank you for your valuable feedback! You may now close this browser window or tab.

A.6 CODEBOOK FOR EMAIL CLASSIFICATION

This appendix contains the codebook that resulted from qualitative analysis of the email communication we conducted with recipients of security or privacy notifications in our study presented in Chapter 6.

Numbers in the % columns are relative to the total number of email conversations about domains with security issues (n = 167) and privacy issues (n = 414), respectively.

Table A.10: Codebook for email classification (Sentiment / Information)

Code	Description	Examples	Counter-examples	Requires	# of Conversations			
					Security		Privacy	
					n	%	n	%
Sentiment								
thanks	The recipient thanks us for the notification.	"Thank you for your notification"			125	74.9	232	56.0
great-project	The recipient expressed that they liked our project.	"Thank you for your work," "We need more projects and people like you!", "good luck with the project," "In case you find any other vulnerabilities I'd be extremely grateful if you would let me know"	"Thank you for your notification," "Many thanks for your two messages, including the valuable advice"		27	16.2	12	2.9
negative	The recipient did not like our project or our notification.	"PISS OFF!!!", "telling a UK business what to do is completely unacceptable," "stop sending threatening emails, it's stupid"	"I would like to be excluded from your project"		3	1.8	22	5.3
More information								
more-info	The recipient asks for more information about, e. g., our project, our checks, about the GDPR, about Git.	"Do you think the GDPR applies to us?", "What needs to be changed?"	"How are you?", "Can you fix this for us?", "Can you exclude us?"		42	25.1	131	31.6
privacy-check	The email contains a question about our privacy checks.	"How does your check recognize a privacy policy?"	"What must be included in a privacy policy?"	more-info	1	0.6	33	8.0
cookie-name	The email contains a question for a cookie name.	"Could you be so kind to specify name of the cookie you are referring to?"		privacy-check	1	0.6	22	5.3

Continued on next page

Table A.11: Codebook for email classification (Information / Actions)

Code	Description	Examples	Counter-examples	Requires	# of Conversations			
					Security		Privacy	
					n	%	n	%
More information (cont.)								
git-check	The email contains a question about our Git check.	"Could you please provide more details about your findings and the actions performed by your automated [Git] analysis?"		more-info	18	10.8	0	0.0
git-url	The email asks for the URL of the Git repository or a file (e. g., config) inside the repository.	"At which URL have you been able to access the repository?"		git-check	9	5.4	0	0.0
project-info	The email contains a question about our research project in general.	"How have you selected our website?", "Is it also possible to trigger this check one way or another?"		more-info	8	4.8	20	4.8
state	The email asks if the issue is still present on the website.	"Could you check again?"		more-info	9	5.4	28	6.8
fix-this-plz	The recipient asks us to fix the issue on their behalf.	"How do you fix this? Can you do that?"		more-info	2	1.2	0	0.0
Performed actions								
fixed	The email states that the issue has (presumably) been fixed.	"This has been resolved," "I've updated my nginx configuration to deny all access to '.' directories," "the security issue should be fixed now"			75	44.9	60	14.5
will-handle	The email states that the recipient will look into the issue or fix the issue in the future.	"I will arrange according to your advice," "You can assume that the website's communication will be encrypted within the next hours," "We will fix it asap"			27	16.2	157	37.9
notified	The recipient notified someone else in order to fix or look into the issue.	"I will get in touch immediately with the person that created our website," "I've forwarded your message to domain owner"	"We will handle this."		17	10.2	66	15.9

Table A.12: Codebook for email classification (Correctness / Language)

.

Code	Description	Examples	Counter-examples	Requires	# of Conversations			
					Security		Privacy	
					n	%	n	%
Correctness								
false-positive	The recipient thinks that the current state of their website is secure / compliant.	"I thought GDPR does not apply to our website," "We do not gather any third party cookie data from visitors"			10	6.0	75	18.1
git-no-sensitive	The recipient thinks that the Git repository does not contain any sensitive data.	"the git repo doesn't contain any confidential information," "the repository is also published at [URL]"			14	8.4	0	0.0
laws-not-apply	The recipient thinks that the privacy laws do not apply to them.	"I thought GDPR does not apply to our website," "Our web page is not public," "We do not process personal data," "We do not have cookies for visitors acceptance and only visitors that subscribe newsletter provide their email"		false-positive	0	0.0	40	9.7
laws-not-in-uk	The recipient thinks that EU privacy laws do not apply to them because they are in the UK.	"Why do you contact an [sic!] UK business?"		laws-not-apply	0	0.0	5	1.2
Language								
expected-german	The recipient asks why we sent emails in English and not German, the language of our institutions' country.	"Why do you send an English email to a German as a German research institute?	"Feel free to also contact me in German"		0	0.0	6	1.4
translate	The recipient asks for a translation into another language (most frequently German).	"If you want to communicate with me, then please write in German!", "In German, please," "is there possibly a 'German version' of this email?"	"Feel free to also contact me in German"		1	0.6	12	2.9

Table A.13: Codebook for email classification (Other)

Code	Description	Examples	Counter-examples	Requires	# of Conversations			
					Security		Privacy	
					n	%	n	%
Other								
unsure-scam	The recipient is unsure if the mail is spam / a scam.	"is this a real email or a phishing attempt," "This looks extremely suspicious to me in its content, tone, and method of delivery," "this looks like spam," "Is this a legitimate email?"	"somehow sounds legitimate"		12	7.2	50	12.1
really-cispa	The recipient is unsure if the email is really from CISPA.			unsure-scam	10	6.0	35	8.5
exclude	The recipient wants to be excluded from our study. Includes conditional exclusion requests.	"Either you call us or I have to ask you to exclude our website"			7	4.2	34	8.2

[1] Yasemin Acar, Michael Backes, Sascha Fahl, Simson Garfinkel, Doowon Kim, Michelle L. Mazurek, and Christian Stransky. "Comparing the Usability of Cryptographic APIs." In: *Proceedings of the 2017 IEEE Symposium on Security and Privacy*. S&P '17. San Jose, CA, USA: IEEE Computer Society, 2017, pp. 154–171. DOI: 10.1109/SP.2017.52. URL: https://ieeexplore.ieee.org/document/7958576.

[2] Yasemin Acar, Christian Stransky, Dominik Wermke, Michelle L. Mazurek, and Sascha Fahl. "Security Developer Studies with GitHub Users: Exploring a Convenience Sample." In: *Proceedings of the Thirteenth Symposium on Usable Privacy and Security*. SOUPS 2017. Santa Clara, CA, USA: USENIX Association, 2017, pp. 81–95. URL: https://www.usenix.org/conference/soups2017/technical-sessions/presentation/acar.

[3] Alessandro Acquisti. "Nudging Privacy: The Behavioral Economics of Personal Information." In: *IEEE Security & Privacy* 7.6 (Dec. 2009), pp. 82–85. DOI: 10.1109/MSP.2009.163. URL: https://ieeexplore.ieee.org/document/5370707.

[4] Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. "Privacy and human behavior in the age of information." In: *Science* 347.6221 (Jan. 2015), pp. 509–514. DOI: 10.1126/science.aaa1465. URL: https://science.sciencemag.org/content/347/6221/509.

[5] Alessandro Acquisti et al. "Nudges for Privacy and Security: Understanding and Assisting Users' Choices Online." In: *ACM Computing Surveys* 50.3 (Aug. 2017). DOI: 10.2139/ssrn.2859227. URL: https://ssrn.com/abstract=2859227.

[6] Alexa Internet, Inc. *The top 500 sites on the web*. 2022. URL: https://www.alexa.com/topsites.

[7] Valentin Allaire. *BlockAdBlock (v3.2.1)*. 2015. URL: https://github.com/sitexw/BlockAdBlock.

[8] Angry Creative AB. *Custom Cookie Message*. Version 2.2.9. 2018. URL: https://wordpress.org/plugins/custom-cookie-message/.

[9] Apple Inc. *App privacy questions requirement starts December 8*. Nov. 2020. URL: https://developer.apple.com/news/?id=em8fm29e.

[10] Ars Technica Staff. *Firefox, Chrome start calling HTTP connections insecure*. Jan. 2017. URL: https://arstechnica.com/information-technology/2017/01/firefox-chrome-start-calling-http-connections-insecure/.

[11] Article 29 Data Protection Working Party. *Opinion 04/2012 on Cookie Consent Exemption*. Tech. rep. 00879/12/EN WP 194. Brussels, Belgium: European Commission, June 2012. URL: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2012/wp194_en.pdf.

[12] Article 29 Data Protection Working Party. *Working Document 02/2013 providing guidance on obtaining consent for cookies*. Tech. rep. 1676/13/EN WP208. Brussels, Belgium: European Commission, Oct. 2013. URL: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2013/wp208_en.pdf.

[13] Article 29 Data Protection Working Party. *Cookie Sweep Combined Analysis – Report*. Tech. rep. 14/EN WP 229. Brussels, Belgium: European Commission, Nov. 2016. URL: https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=640605.

[14] Article 29 Data Protection Working Party. *Guidelines on Consent under Regulation 2016/679*. Tech. rep. 17/EN WP259 rev.01. Brussels, Belgium: European Commission, Oct. 2018. URL: https://ec.europa.eu/newsroom/article29/items/623051/en.

[15] Rebecca Balebako, Abigail Marsh, Jialiu Lin, Jason Hong, and Lorrie Faith Cranor. "The Privacy and Security Behaviors of Smartphone App Developers." In: *Workshop on Usable Security*. USEC 2014. San Diego, CA, USA: Internet Society, 2014. DOI: 10.14722/usec.2014.23006. URL: https://www.ndss-symposium.org/ndss2014/workshop-usable-security-usec-2014-programme/privacy-and-security-behaviors-smartphone-app-developers/.

[16] Robert Bateman. *CCPA: Does Using Third-Party Cookies Count as Selling Personal Information?* May 2022. URL: https://www.termsfeed.com/blog/ccpa-third-party-cookies-selling-personal-information/#Valuable_Consideration.

[17] Stephanie Bodoni. *Facebook's Like Button Makes Websites Liable, Top EU Court Rules*. July 2019. URL: https://www.bloomberg.com/news/articles/2019-07-29/facebook-s-like-button-makes-websites-liable-top-eu-court-rules.

[18] Sophie C. Boerman, Sanne Kruikemeier, and Frederik J. Zuiderveen Borgesius. "Exploring Motivations for Online Privacy Protection Behavior: Insights From Panel Data." In: *Communication Research* 48.7 (Oct. 2021), pp. 953–977. DOI: 10.1177/0093650218800915. URL: https://journals.sagepub.com/doi/10.1177/0093650218800915.

[19] Dino Bollinger, Karel Kubicek, Carlos Cotrini, and David Basin. "Automating Cookie Consent and GDPR Violation Detection." In: *Proceedings of the 31st USENIX Security Symposium*. USENIX Security '22. Boston, MA, USA: USENIX Association, 2022, pp. 2893–2910. URL: https://www.usenix.org/conference/usenixsecurity22/presentation/bollinger.

[20] Bristows LLC. *Status of implementation of the amendment to Article 5.3 of Directive 2002/58/EC (the "EU Cookie Law")*. Tech. rep. London, United Kingdom: Bristows LLC, June 2015. URL: http://backstrom.fi/wp-content/uploads/2015/06/European-Cookie-Law-Implementation-Survey-June-2015.pdf.

[21] Brontobytes. *Cookie Bar*. 2018. URL: https://wordpress.org/plugins/cookie-bar/.

[22] Jennifer Bryant. *Belgian DPA fines IAB Europe 250K euros over consent framework GDPR violations*. Feb. 2022. URL: https://iapp.org/news/a/belgian-dpa-fines-iab-europe-250k-euros-over-consent-framework-gdpr-violations/.

[23] Matthew Bryant. *TheInternetBackup – Crowdsourced Internet Domain Database*. Jan. 2020. URL: https://web.archive.org/web/20200110214540/https://theinternetbackup.com//#about.

[24] BuiltWith. *CAPTCHA Usage Distribution on the Entire Internet*. Sept. 2021. URL: https://trends.builtwith.com/widgets/captcha/traffic/Entire-Internet.

[25] Matt Burgess. *The tyranny of GDPR popups and the websites failing to adapt*. Aug. 2018. URL: https://www.wired.co.uk/article/gdpr-cookies-eprivacy-regulation-popups.

[26] Matt Burgess. *Europe's Move Against Google Analytics Is Just the Beginning*. Jan. 2022. URL: https://www.wired.com/story/google-analytics-europe-austria-privacy-shield/.

[27] Karoline Busse, Mohammad Tahaei, Katharina Krombholz, Emanuel von Zezschwitz, Matthew Smith, Jing Tian, and Wenyuan Xu. "Cash, Cards or Cryptocurrencies? A Study of Payment Culture in Four Countries." In: *Proceedings of the 5th European Workshop on Usable Security*. EuroUSEC 2020. Virtual Event, Italy: University of Chicago, 2020. URL: https://eusec20.cs.uchicago.edu/eusec20-Busse.pdf.

[28] Virginio Cantoni, Marco Porta, Stefania Ricotti, and Francesca Zanin. "Banner positioning in the masthead area of online newspapers: an eye tracking study." In: *14th International Conference on Computer Systems and Technologies*. CompSysTech '13. Ruse, Bulgaria: ACM, 2013, pp. 145–152. DOI: 10.1145/2516775.2516789. URL: https://dl.acm.org/citation.cfm?id=2516789.

[29] Mike Carter. *Cookie Control*. Version 1.7-1.6. 2018. URL: https://www.drupal.org/project/cookiecontrol.

[30] Amber Case. *The Current State of Micropayments and Web Monetization*. July 2021. URL: https://caseorganic.medium.com/part-ii-the-current-state-of-micropayments-and-web-monetization-50ee2e58d332.

[31] Amber Case. *Who killed the micropayment? A history.* Mar. 2021. URL: https://caseorganic.medium.com/who-killed-the-micropayment-a-history-ec9e6eb39d05.

[32] Catapult Themes. *Cookie Consent*. Version Version 2.3.11. 2018. URL: https://catapultthemes.com/cookie-consent/.

[33] Orçun Çetin, Carlos Gañán, Maciej Korczyński, and Michel van Eeten. "Understanding the role of sender reputation in abuse reporting and cleanup." In: *Journal of Cybersecurity* 2.1 (Dec. 2016), pp. 83–98. DOI: 10.1093/cybsec/tyw005. URL: https://academic.oup.com/cybersecurity/article/2/1/83/2629556.

[34] Orçun Çetin, Carlos Gañán, Maciej Korczyński, and Michel van Eeten. "Make Notifications Great Again: Learning How to Notify in the Age of Large-Scale Vulnerability Scanning." In: *16th Annual Workshop on the Economics of Information Security*. WEIS 2017. La Jolla, CA, USA, 2017. URL: https://weis2017.econinfosec.org/wp-content/uploads/sites/3/2017/05/WEIS_2017_paper_17.pdf.

[35] Civic. *Cookie Control by CIVIC – GDPR Cookie Compliance Solution*. 2018. URL: https://www.civicuk.com/cookie-control.

[36] Victoria Clarke and Virginia Braun. "Thematic Analysis." In: *Encyclopedia of Critical Psychology*. Ed. by Thomas Teo. New York, NY, USA: Springer New York, 2014, pp. 1947–1952. DOI: 10.1007/978-1-4614-5583-7_311. URL: https://link.springer.com/referenceworkentry/10.1007/978-1-4614-5583-7_311.

[37] James Clayton. *Europe agrees new law to curb Big Tech dominance*. Mar. 2022. URL: https://www.bbc.com/news/technology-60870287.

[38] Clickio. *Clickio GDPR Consent Tool*. 2018. URL: http://gdpr.clickio.com/.

[39] Cliqz Privacy Team. *Do you consent?* June 2018. URL: https://whotracks.me/blog/update_jun_2018.html.

[40] CMS Law.Tax. *GDPR Enforcement Tracker*. 2022. URL: https://www.enforcementtracker.com/.

[41] Keith Collins. *How one programmer broke the internet by deleting a tiny piece of code*. Mar. 2016. URL: https://qz.com/646467/how-one-programmer-broke-the-internet-by-deleting-a-tiny-piece-of-code/.

[42] Cookie Information A/S. *Cookie Information*. 2018. URL: https://cookieinformation.com/.

[43] Forbrukerrådet (Norwegian Consumer Council). *Deceived by Design – How tech companies use dark patterns to discourage us from exercising our rights to privacy*. Tech. rep. Oslo, Norway: Forbrukerrådet, June 2018. URL: https://fil.forbrukerradet.no/wp-content/uploads/2018/06/2018-06-27-deceived-by-design-final.pdf.

[44] Lorrie Cranor. "Necessary But Not Sufficient: Standardized Mechanisms for Privacy Notice and Choice." In: *Journal on Telecommunications and High Technology Law* 10 (2012), pp. 273–307. URL: http://jthtl.org/content/articles/V10I2/JTHTLv10i2_Cranor.PDF.

[45] Lorrie Cranor, Marc Langheinrich, Massimo Marchiori, Martin Presler-Marshall, and Joseph Reagle. *The Platform for Privacy Preferences 1.0 (P3P1.0) Specification*. Tech. rep. Cambridge, MA, USA: W3C, Apr. 2002. URL: https://www.w3.org/TR/P3P/.

[46] Dave Crocker. *Mailbox Names for Common Services, Roles and Functions*. Tech. rep. RFC 2142. Fremont, CA, USA: Internet Engineering Task Force, May 1997. URL: https://www.ietf.org/rfc/rfc2142.txt.

[47] Cybot. *Cookiebot – GDPR and ePrivacy compliant online tracking*. 2018. URL: https://www.cookiebot.com/.

[48] Adrian Dabrowski, Georg Merzdovnik, Johanna Ullrich, Gerald Sendera, and Edgar Weippl. "Measuring Cookies and Privacy in a Post-GDPR World." In: *Proceedings of the 20th International Conference on Passive and Active Measurement*. PAM 2019. Puerto Varas, Chile: Springer Nature Switzerland AG, 2019, pp. 258–270. DOI: 10.1007/978-3-030-15986-3_17. URL: https://link.springer.com/chapter/10.1007/978-3-030-15986-3_17.

[49] Anastasia Danilova, Alena Naiakshina, Johanna Deuter, and Matthew Smith. "Replication: On the Ecological Validity of Online Security Developer Studies: Exploring Deception in a Password-Storage Study with Freelancers." In: *Proceedings of the Sixteenth Symposium on Usable Privacy and Security*. SOUPS 2020. Virtual Event, USA: USENIX Association, 2020, pp. 165–183. URL: https://www.usenix.org/system/files/soups2020-danilova.pdf.

[50] Data USA. *Web Developers*. 2022. URL: https://datausa.io/profile/soc/web-developers.

[51] Kevin Davis. *You (probably) don't need ReCAPTCHA*. June 2019. URL: https://kevv.net/you-probably-dont-need-recaptcha/.

[52] Martin Degeling, Christine Utz, Christopher Lentzsch, Henry Hosseini, Florian Schaub, and Thorsten Holz. "We Value Your Privacy ... Now Take Some Cookies: Measuring the GDPR's Impact on Web Privacy." In: *26th Annual Network and Distributed System Security Symposium*. NDSS '19. San Diego, CA, USA: Internet Society, 2019. DOI: 10.14722/ndss.2019.23378. URL: https://www.ndss-symposium.org/ndss-paper/we-value-your-privacy-now-take-some-cookies-measuring-the-gdprs-impact-on-web-privacy/.

[53] dFactory. *Cookie Notice for GDPR*. Version 1.2.45. 2018. URL: https://dfactory.eu/products/cookie-notice/.

[54] Didomi. *Privacy Center: Build trust with your customers*. 2018. URL: https://www.didomi.io/en/privacy-center.

[55] Digital Advertising Alliance. *YourAdChoices*. June 2022. URL: https://youradchoices.com/.

[56] DuckDuckGo. *DuckDuckGo Tracker Radar*. Feb. 2021. URL: https://github.com/duckduckgo/tracker-radar/.

[57] Zakir Durumeric. *Cached Chrome Top Million Websites*. Dec. 2022. URL: https://github.com/zakird/crux-top-lists.

[58] Zakir Durumeric, Eric Wustrow, and J. Alex Halderman. "ZMap: Fast Internet-wide Scanning and Its Security Applications." In: *Proceedings of the 22nd USENIX Security Symposium*. USENIX Security '13. Washington, DC, USA: USENIX Association, 2013, pp. 605–619. URL: https://www.usenix.org/system/files/conference/usenixsecurity13/sec13-paper_durumeric.pdf.

[59] Zakir Durumeric et al. "The Matter of Heartbleed." In: *Proceedings of the 2014 ACM Internet Measurement Conference*. IMC '14. Vancouver, BC, Canada: ACM, 2014, pp. 475–488. DOI: 10.1145/2663716.2663755. URL: https://dl.acm.org/doi/10.1145/2663716.2663755.

[60] The EasyList authors. *EasyList*. 2022. URL: https://easylist.to/easylist/easylist.txt.

[61] The EasyList authors. *EasyList Cookie List*. 2022. URL: https://github.com/easylist/easylist/tree/master/easylist_cookie.

[62] Serge Egelman, Lorrie Faith Cranor, and Jason Hong. "You've Been Warned: An Empirical Study of the Effectiveness of Web Browser Phishing Warnings." In: *Proceedings of the 26th Annual CHI Conference on Human Factors in Computing Systems*. CHI '08. Florence, Italy: ACM, 2008, pp. 1065–1074. DOI: 10.1145/1357054.1357219. URL: https://dl.acm.org/doi/10.1145/1357054.1357219.

[63] Steven Englehardt and Arvind Narayanan. "Online Tracking: A 1-million-site Measurement and Analysis." In: *Proceedings of the 2016 ACM Conference on Computer and Communications Security*. CCS '16. Vienna, Austria: ACM, 2016, pp. 1388–1401. DOI: 10.1145/2976749.2978313. URL: https://dl.acm.org/doi/10.1145/2976749.2978313.

[64] Steven Englehardt, Dillon Reisman, Christian Eubank, Peter Zimmerman, Jonathan Mayer, Arvind Narayanan, and Edward W. Felten. "Cookies That Give You Away: The Surveillance Implications of Web Tracking." In: *Proceedings of the 24th International Conference on World Wide Web*. WWW '15. Florence, Italy: IW3C3, 2015, pp. 289–299. DOI: 10.1145/2736277.2741679. URL: https://dl.acm.org/doi/10.1145/2736277.2741679.

[65] EthicalAds. *Our Advertising Vision*. 2022. URL: https://www.ethicalads.io/advertising-vision/.

[66] European Commission. *Proposal for a Regulation on Privacy and Electronic Communications*. Jan. 2017. URL: https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-privacy-and-electronic-communications.

[67] European Commission. *What is personal data?* 2022. URL: https://ec.europa.eu/info/law/law-topic/data-protection/reform/what-personal-data_en.

[68] European Court of Justice. *Judgment of the Court of 6 October 2015 in Case C-362/14 – Schrems I*. Oct. 2015. URL: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A62014CJ0362.

[69] European Court of Justice. *Judgment of the Court of 1 October 2019 in Case C-673/17 – Planet49*. Oct. 2019. URL: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:62017CJ0673.

[70] European Court of Justice. *Judgment of the Court of 16 July 2020 in Case C-311/18 – Schrems II*. Oct. 2020. URL: https://eur-lex.europa.eu/legal-content/en/ALL/?uri=CELEX:62018CJ0311.

[71] European Data Protection Board. *Guidelines 05/2020 on consent under Regulation 2016/679*. Tech. rep. Version 1.1. Brussels, Belgium: European Data Protection Board, May 2018. URL: https://edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_202005_consent_en.pdf.

[72] European Data Protection Board. *Opinion 5/2019 on the interplay between the ePrivacy Directive and the GDPR, in particular regarding the competence, tasks and powers of data protection authorities*. Tech. rep. Mar. 2019. URL: https://edpb.europa.eu/sites/edpb/files/files/file1/201905_edpb_opinion_eprivacydir_gdpr_interplay_en_0.pdf.

[73] European Digital Rights. *Ethical Web Dev – Guide for Ethical Website Development and Maintenance*. Jan. 2020. URL: https://edri.org/files/ethical_web_dev_web.pdf.

[74] European Free Trade Association. *General Data Protection Regulation incorporated into the EEA Agreement*. July 2018. URL: https://www.efta.int/EEA/news/General-Data-Protection-Regulation-incorporated-EEA-Agreement-509291.

[75] The European Parliament and the Council of the European Union. *Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data*. Official Journal of the European Communities. Oct. 1995. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A31995L0046.

[76] The European Parliament and the Council of the European Union. *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*. Official Journal of the European Communities. July 2002. URL: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32002L0058.

[77] The European Parliament and the Council of the European Union. *Directive 2009/136/EC of the European Parliament and of the Council of 25 November 2009 amending Directive 2002/22/EC, Directive 2002/58/EC and Regulation (EC) No 2006/2004*. Official Journal of the European

Union, L 337/11. Nov. 2009. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32009L0136.

[78] The European Parliament and the 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)*. Official Journal of the European Union, L 119/1. Apr. 2016. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

[79] Evidon Inc. *Evidon – Universal Consent Platform for Site, App, and Ad Compliance*. 2018. URL: https://www.evidon.com/solutions/universal-consent/.

[80] Todd Feathers, Simon Fondrie-Teitler, Angie Waller, and Surya Mattu. *Facebook Is Receiving Sensitive Medical Information from Hospital Websites*. June 2022. URL: https://themarkup.org/pixel-hunt/2022/06/16/facebook-is-receiving-sensitive-medical-information-from-hospital-websites.

[81] Adrienne Porter Felt, Alex Ainslie, Robert W. Reeder, Sunny Consolvo, Somas Thyagaraja, Alan Bettes, Helen Harris, and Jeff Grimes. "Improving SSL Warnings: Comprehension and Adherence." In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. CHI '15. Seoul, Republic of Korea: ACM, 2015, pp. 2893–2902. DOI: 10.1145/2702123.2702442. URL: https://dl.acm.org/doi/10.1145/2702123.2702442.

[82] Adrienne Porter Felt, Richard Barnes, April King, Chris Palmer, Chris Bentzel, and Parisa Tabriz. "Measuring HTTPS Adoption on the Web." In: *Proceedings of the 26th USENIX Security Symposium*. USENIX Security '17. Vancouver, BC, Canada: USENIX Association, 2017, pp. 1323–1338. URL: https://www.usenix.org/conference/usenixsecurity17/technical-sessions/presentation/felt.

[83] Roy T. Fielding and David Singer. *Tracking Preference Expression (DNT)*. Tech. rep. Cambridge, MA, USA: W3C, Jan. 2019. URL: https://www.w3.org/TR/2019/NOTE-tracking-dnt-20190117/.

[84] Ronald A. Fisher. "On the Interpretation of χ^2 from Contingency Tables, and the Calculation of P." In: *Journal of the Royal Statistical Society* 85.1 (Jan. 1922), pp. 87–94. DOI: 10.2307/2340521. URL: https://www.jstor.org/stable/2340521.

[85] Ronald A. Fisher. "The Logic of Inductive Interference." In: *Journal of the Royal Statistical Society* 98.1 (1935), pp. 39–82. DOI: 10.2307/2342435. URL: https://www.jstor.org/stable/2342435.

[86] Edwin Foudil and Yakov Shafranovich. *security.txt – A proposed standard which allows websites to define security policies*. 2017. URL: https://securitytxt.org/.

[87] Edwin Foudil and Yakov Shafranovich. *A File Format to Aid in Security Vulnerability Disclosure*. Tech. rep. RFC 9116. Fremont, CA, USA, 2022. URL: https://www.rfc-editor.org/rfc/rfc9116.

[88] Suzanne Frey. *New safety section in Google Play will give transparency into how apps use data*. May 2021. URL: https://android-developers.googleblog.com/2021/05/new-safety-section-in-google-play-will.html.

[89] Vitaly Friedman. *Privacy UX: Better Cookie Consent Experiences*. Apr. 2019. URL: https://www.smashingmagazine.com/2019/04/privacy-ux-better-cookie-consent-experiences/.

[90] Stacia Garlach and Daniel Suthers. "'I'm supposed to see that?' AdChoices Usability in the Mobile Environment." In: *Proceedings of the 51st Hawaii International Conference on System Sciences*. HICSS 2018. Waikoloa Village, HI, USA: University of Hawai'i at Mānoa, 2018, pp. 3779–3788. DOI: 10.24251/hicss.2018.476. URL: http://hdl.handle.net/10125/50364.

[91] Ghostery GmbH. *WhoTracks.me*. May 2022. URL: https://github.com/whotracksme/whotracks.me.

[92] Ghostery GmbH. *WhoTracks.me – Trackers*. July 2022. URL: https://whotracks.me/trackers.html.

[93] GitHub, Inc. *GitHub Terms of Service*. Nov. 2020. URL: https://docs.github.com/en/site-policy/github-terms/github-terms-of-service.

[94] GitHub, Inc. *GitHub Privacy Statement*. Sept. 2022. URL: https://docs.github.com/en/site-policy/privacy-policies/github-privacy-statement.

[95] GitHub, Inc. *Setting your commit email address*. 2022. URL: https://docs.github.com/en/account-and-profile/setting-up-and-managing-your-personal-account-on-github/managing-email-preferences/setting-your-commit-email-address.

[96] The Global Privacy Control group. *Global Privacy Control*. Oct. 2020. URL: https://globalprivacycontrol.org/.

[97] Cody Godwin. *US lawmakers introduce bills targeting Big Tech*. June 2021. URL: https://www.bbc.com/news/technology-57450345.

[98] Google, Inc. *Privacy controls in Google Analytics*. 2021. URL: https://support.google.com/analytics/answer/9019185.

[99] Google, Inc. *CrUX API*. 2022. URL: https://developer.chrome.com/docs/crux/api/.

[100] Google, Inc. *Embed videos and playlists*. 2022. URL: https://support.google.com/youtube/answer/171780?hl=en-GB.

[101] Google, Inc. *IP Anonymization (or IP masking) in Universal Analytics*. 2022. URL: https://support.google.com/analytics/answer/2763052?hl=en.

[102] Peter Leo Gorski, Luigi Lo Iacono, Dominik Wermke, Christian Stansky, Sebastian Möller, Yasemin Acar, and Sascha Fahl. "Developers Deserve Security Warnings, Too: On the Effect of Integrated Security Advice on Cryptographic API Misuse." In: *Fourteenth Symposium on Usable Privacy and Security*. SOUPS 2018. Baltimore, MD, USA: USENIX Association, 2018, pp. 265–280. DOI: 10.5555/3291228.3291250. URL: https://dl.acm.org/doi/10.5555/3291228.3291250.

[103] Björn Greif. *Study: Google Is the Biggest Beneficiary of the GDPR*. Oct. 2018. URL: https://cliqz.com/en/magazine/study-google-is-the-biggest-beneficiary-of-the-gdpr.

[104] Sacha Greif. *The State of JS 2021*. 2022. URL: https://2021.stateofjs.com/en-US/demographics.

[105] Martin Gundersen. *Uncovering the Disqus Data Machine*. Dec. 2019. URL: https://twitter.com/martingund/status/1207327648093003777.

[106] Vicki Ha, Kori Inkpen, Farah Al Shaar, and Lina Hdeib. "An Examination of User Perception and Misconception of Internet Cookies." In: *CHI '06 Extended Abstracts on Human Factors in Computing Systems*. CHI EA '06. Montréal, QC, Canada: ACM, 2006, pp. 833–838. DOI: 10.1145/1125451.1125615. URL: https://dl.acm.org/doi/10.1145/1125451.1125615.

[107] Hana Habib, Yixin Zou, Aditi Jannu, Neha Sridhar, Chelse Swoopes, Alessandro Acquisti, Lorrie Faith Cranor, Norman Sadeh, and Florian Schaub. "An Empirical Analysis of Data Deletion and Opt-Out Choices on 150 Websites." In: *Fifteenth Symposium On Usable Privacy and Security*. SOUPS 2019. Santa Clara, CA, USA: USENIX Association, 2019, pp. 387–406. URL: https://www.usenix.org/conference/soups2019/presentation/habib.

[108] Irit Hadar, Tomer Hasson, Oshrat Ayalon, Eran Toch, Michael Birnhack, Sofia Sherman, and Arod Balissa. "Privacy by designers: software developers' privacy mindset." In: *Empirical Software Engineering* 23.1 (Feb. 2018), pp. 259–289. DOI: 10.1007/s10664-017-9517-1. URL: https://link.springer.com/article/10.1007/s10664-017-9517-1.

[109] Catherine Han, Irwin Reyes, Amit Elazari Bar On, Joel Reardon, Álvaro Feal, Kenneth A. Bamberger, Serge Egelman, and Narseo Vallina-Rodriguez. "Do You Get What You Pay For? Comparing the Privacy Behaviors of Free vs. Paid Apps." In: *3rd Workshop on Technology and Consumer Protection*. ConPro '19. San Francisco, CA, USA: IEEE Computer Society, 2019. URL: https://www.ieee-security.org/TC/SPW2019/ConPro/papers/han-conpro19.pdf.

[110] Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G. Shin, and Karl Aberer. "Polisis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning." In: *Proceedings of the 27th USENIX Security Symposium*. USENIX Security '18. Baltimore, MD, USA: USENIX Association, 2018, pp. 531–548. URL: https://www.

usenix.org/conference/usenixsecurity18/presentation/harkous.

[111] Ilse Heine. *3 Years Later: An Analysis of GDPR Enforcement*. Sept. 2021. URL: https://www.csis.org/blogs/strategic-technologies-blog/3-years-later-analysis-gdpr-enforcement.

[112] James Hercher. *Google Analytics To Stop Logging IP Addresses And Sunset Old Versions In Privacy Standards Overhaul*. Mar. 2022. URL: https://www.adexchanger.com/online-advertising/google-analytics-to-stop-logging-ip-addresses-and-sunset-old-versions-in-privacy-standards-overhaul/.

[113] Alex Hern and Jim Waterson. *Sites block users, shut down activities and flood inboxes as GDPR rules loom*. May 2018. URL: https://www.theguardian.com/technology/2018/may/24/sites-block-eu-users-before-gdpr-takes-effect.

[114] Maximilian Hils, Daniel W. Woods, and Rainer Böhme. "Measuring the Emergence of Consent Management on the Web." In: *Proceedings of the 2020 ACM Internet Measurement Conference*. IMC '20. Virtual Event, USA: ACM, 2020, pp. 317–332. DOI: 10.1145/3419394.3423647. URL: https://dl.acm.org/doi/10.1145/3419394.3423647.

[115] Sture Holm. "A Simple Sequentially Rejective Multiple Test Procedure." In: *Scandinavian Journal of Statistics* 6.2 (1979), pp. 65–70. URL: https://www.jstor.org/stable/4615733.

[116] Amanda Holpuch. *'You are the worst human being': man confronts Fox News host Tucker Carlson*. July 2021. URL: https://www.theguardian.com/media/2021/jul/25/tucker-carlson-fox-news-worst-human-being-viral-video-montana.

[117] Henry Hosseini, Martin Degeling, Christine Utz, and Thomas Hupperich. "Unifying Privacy Policy Detection." In: *Proceedings on Privacy Enhancing Technologies* 2021.4 (July 2021), pp. 480–499. DOI: 10.2478/popets-2021-0081. URL: https://petsymposium.org/popets/2021/popets-2021-0081.php.

[118] Xuehui Hu and Nishanth Sastry. "What a Tangled Web We Weave: Understanding the Interconnectedness of the Third Party Cookie Ecosystem." In: *Proceedings of the 12th ACM Conference on Web Science*. WebSci '20. Southampton, United Kingdom: ACM, 2020, pp. 76–85. DOI: 10.1145/3394231.3397897. URL: https://dl.acm.org/doi/10.1145/3394231.3397897.

[119] Xuehui Hu, Nishanth Sastry, and Mainack Mondal. "CCCC: Corralling Cookies into Categories with CookieMonster." In: *Proceedings of the 13th ACM Web Science Conference*. WebSci '21. Virtual Event, United Kingdom: ACM, 2021, pp. 234–242. DOI: 10.1145/3447535.3462509. URL: https://dl.acm.org/doi/10.1145/3447535.3462509.

[120] Xuehui (Rachel) Hu, Guillermo Suaréz de Tangil, and Nishanth Sastry. "Multi-country Study of Third Party Trackers from Real Browser Histories." In: *Proceedings of the 5th IEEE European Symposium on Security and Privacy*. EuroS&P 2020. Virtual Event, Italy: IEEE Computer Society, 2020, pp. 70–86. DOI: 10.1109/EuroSP48549.2020.00013. URL: https://ieeexplore.ieee.org/document/9230391.

[121] Patrick Hulce. *The Web Almanac – Third Parties*. Nov. 2019. URL: https://almanac.httparchive.org/en/2019/third-parties.

[122] Patrick Hulce. *Third Party Web*. Oct. 2022. URL: https://github.com/patrickhulce/third-party-web/.

[123] Soheil Human, Harshvardhan J. Pandit, Victor Pierre Morel, Cristiana Santos, Martin Degeling, Arianna Rossi, Wilhelmina Botes, Vitor Jesus, and Irene Kamara. "Data Protection and Consenting Communication Mechanisms: Current Open Proposals and Challenges." In: *2022 International Workshop on Privacy Engineering*. IWPE '22. Genoa, Italy: Vienna University of Economics and Business, 2022. URL: https://epub.wu.ac.at/8604/.

[124] Soheil Human, Max Schrems, Alan Toner, Gerben, and Ben Wagner. *Advanced Data Protection Control (ADPC)*. Tech. rep. Vienna, Austria: Vienna University of Economics and Business, Oct. 2021. URL: https://www.dataprotectioncontrol.org/spec/.

[125] Tatum Hunter. *You scheduled an abortion. Planned Parenthood's website could tell Facebook.* June 2022. URL: https://www.washingtonpost.com/technology/2022/06/29/planned-parenthood-privacy/.

[126] IAB Europe. *Vendor List*. Version 124. Dec. 2018. URL: https://web.archive.org/web/20181213024456/https://vendorlist.consensu.org/vendorlist.json.

[127] IAB Europe. *CMP List*. 2022. URL: https://iabeurope.eu/cmp-list/.

[128] IAB Europe. *Vendor List TCF v2.0*. 2022. URL: https://iabeurope.eu/vendor-list-tcf-v2-0/.

[129] Muhammad Ikram, Rahat Masood, Gareth Tyson, Mohamed Ali Kaafar, Noha Loizon, and Roya Ensafi. "The Chain of Implicit Trust: An Analysis of the Web Third-party Resources Loading." In: *The Web Conference 2019 – Proceedings of The World Wide Web Conference*. WWW '19. San Francisco, CA, USA: IW3C2, 2019, pp. 2851–2857. DOI: 10.1145/3308558.3313521. URL: https://dl.acm.org/doi/10.1145/3308558.3313521.

[130] Information Commissioner's Office. *Protecting personal data in online services: learning from the mistakes of others*. Tech. rep. Wilmslow, United Kingdom: Information Commissioner's Office, May 2014. URL: https://ico.org.uk/media/for-organisations/documents/1042221/protecting-personal-data-in-online-services-learning-from-the-mistakes-of-others.pdf.

[131] Interactive Advertising Bureau. *IAB CCPA Compliance Framework for Publishers & Technology Companies*. Tech. rep. New York, NY, USA: Interactive Advertising Bureau, 2021. URL: https://iabtechlab.com/standards/ccpa/.

[132] Interactive Advertising Bureau Europe. *GDPR Transparency and Consent Framework*. Tech. rep. Brussels, Belgium: Interactive Advertising Bureau Europe, 2019. URL: https://iabtechlab.com/standards/gdpr-transparency-and-consent-framework/.

[133] Interledger Foundation. *Interledger: Open and Inclusive Payments*. May 2022. URL: https://interledger.org/.

[134] The Internet Archive. *Wayback Machine*. Dec. 2022. URL: https://web.archive.org/.

[135] Irish Data Protection Commission. *Guidance note: Cookies and other tracking technologies*. Tech. rep. Dublin, Ireland: Data Protection Commission, Apr. 2020. URL: https://www.dataprotection.ie/sites/default/files/uploads/2020-04/Guidance%20note%20on%20cookies%20and%20other%20tracking%20technologies.pdf.

[136] Shubham Jain and Janne Lindqvist. "Should I Protect You? Understanding Developers' Behavior to Privacy-Preserving APIs." In: *Workshop on Usable Security*. USEC 2014. San Diego, CA, USA: Internet Society, 2014. DOI: 10.14722/usec.2014.23045. URL: https://www.ndss-symposium.org/ndss2014/workshop-usable-security-usec-2014-programme/should-i-protect-you-understanding-developers-behavior-privacy-preserving-apis/.

[137] Miroslav Jakúbek and Stefan Panic. *Austria: "Cookie Walls / Paywalls" hybrids are permissible?* Dec. 2018. URL: https://www.lexology.com/library/detail.aspx?g=75a50269-b0fa-471d-932b-dce5fe2e0c66.

[138] Jaohawi AB. *Consent Manager Provider (CMP)*. 2018. URL: https://www.consentmanager.net/.

[139] Arjaldo Karaj, Sam Macbeth, Rémi Berson, and Josep M. Pujol. *WhoTracks.Me: Monitoring the online tracking landscape at scale*. 2018. DOI: 10.48550/arXiv.1804.08959. arXiv: 1804.08959. URL: https://arxiv.org/abs/1804.08959v1.

[140] Arjaldo Karaj, Sam Macbeth, Rémi Berson, and Josep M. Pujol. *WhoTracks.Me: Shedding light on the opaque world of online tracking*. 2019. DOI: 10.48550/arXiv.1804.08959. arXiv: 1804.08959. URL: https://arxiv.org/abs/1804.08959v2.

[141] Farzaneh Karegar, Nina Gerber, Melanie Volkamer, and Simone Fischer-Hübner. "Helping John to Make Informed Decisions on Using Social Login." In: *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*. SAC '18. Pau, France: ACM, 2018, pp. 1165–1174. DOI: 10.1145/3167132.3167259. URL: https://dl.acm.org/doi/10.1145/3167132.3167259.

[142] Jon Keegan and Dara Kerr. *Online Abortion Pill Provider Hey Jane Used Tracking Tools That Sent Visitor Data to Meta, Google, and Others*. July 2022. URL: https://themarkup.org/pixel-hunt/2022/07/01/online-abortion-pill-provider-hey-jane-used-tracking-tools-that-sent-visitor-data-to-meta-google-and-others.

[143] Daniel Kladnik. *I don't care about cookies*. Version 3.0.0. 2019. URL: https://www.i-dont-care-about-cookies.eu/.

[144] Konrad Kollnig, Reuben Binns, Pierre Dewitte, Max Van Kleek, Ge Wang, Daniel Omeiza, Helena Webb, and Nigel Shadbolt. "A Fait Accompli? An Empirical Study into the Absence of Consent to Third-Party Tracking in Android Apps." In: *Proceedings of the Seventeenth Symposium on Usable Privacy and Security*. SOUPS 2021. Virtual Event, USA: USENIX Association, 2021, pp. 181–195. URL: https://www.usenix.org/system/files/soups2021-kollnig.pdf.

[145] Konferenz der unabhängigen Datenschutzaufsichtsbehörden des Bundes und der Länder. *Hinweise zum Einsatz von Google Analytics im nicht-öffentlichen Bereich (in German; Notes concerning the use of Google Analytics in the non-public sector)*. Tech. rep. Ansbach, Germany: Datenschutzkonferenz (DSK), May 2020. URL: https://www.datenschutzkonferenz-online.de/media/dskb/20200526_beschluss_hinweise_zum_einsatz_von_google_analytics.pdf.

[146] Marc Kührer, Thomas Hupperich, Christian Rossow, and Thorsten Holz. "Exit from Hell? Reducing the Impact of Amplification DDoS Attacks." In: *Proceedings of the 23rd USENIX Security Symposium*. USENIX Security '14. San Diego, CA, USA: USENIX Association, 2014, pp. 111–125. URL: https://www.usenix.org/conference/usenixsecurity14/technical-sessions/presentation/kuhrer.

[147] Oksana Kulyk, Annika Hilt, Nina Gerber, and Melanie Volkamer. "'This Website Uses Cookies': Users' Perceptions and Reactions to the Cookie Disclaimer." In: *3rd European Workshop on Usable Security*. EuroUSEC 2018. London, United Kingdom: Internet Society, 2018. DOI: 10.14722/eurousec.2018.23012. URL: https://www.ndss-symposium.org/wp-content/uploads/2018/06/eurousec2018_12_Kulyk_paper.pdf.

[148] Oksana Kulyk, Peter Mayer, Oliver Käfer, and Melanie Volkamer. "A Concept and Evaluation of Usable and Fine-Grained Privacy-Friendly Cookie Settings Interface." In: *Proceedings of the 17th IEEE International Conference on Trust, Security and Privacy in Computing and Communications*. TrustCom 2018. New York, NY, USA: IEEE Computer Society, 2018. DOI: 10.1109/TrustCom/BigDataSE.2018.00148. URL: https://ieeexplore.ieee.org/document/8456017.

[149] Ravie Lakshmanan. *German Court Rules Websites Embedding Google Fonts Violates GDPR*. Jan. 2022. URL: https://thehackernews.com/2022/01/german-court-rules-websites-embedding.html.

[150] Landgericht München. *Urteil 3 O 17493/20 vom 19.01.2022*. Jan. 2022. URL: https://rewis.io/urteile/urteil/lhm-20-01-2022-3-o-1749320/.

[151] Tobias Lauinger, Abdelberi Chaabane, Sajjad Arshad, William Robertson, Christo Wilson, and Engin Kirda. "Thou Shalt Not Depend on Me: Analysing the Use of Outdated JavaScript Libraries on the Web." In: *Proceedings of the 2017 Network and Distributed System Security Symposium*. NDSS '17. San Diego, CA, USA: Internet Society, 2017. DOI: 10.14722/ndss.2017.2341. URL: https://www.ndss-symposium.org/ndss2017/ndss-2017-programme/thou-shalt-not-depend-me-analysing-use-outdated-javascript-libraries-web/.

[152] Victor Le Pochat, Tom Van Goethem, Samaneh Talajizadehkhoob, Maciej Korczyński, and Wouter Joosen. "TRANCO: A Research-Oriented Top Sites Ranking Hardened Against Manipulation." In: *Proceedings of the 26th Annual Network and Distributed System Security Symposium*. NDSS '19. San Diego, CA, USA: Internet Society, 2019. DOI: 10.14722/ndss.2019.23386. URL: https://www.ndss-symposium.org/wp-content/uploads/2019/02/ndss2019_01B-3_LePochat_paper.pdf.

[153] Jason Lemieux. *Disqus: Is Your Data Worth Trading for Convenience?* Mar. 2017. URL: https://replyable.com/2017/03/disqus-is-your-data-worth-trading-for-convenience/.

[154] Pedro Leon, Blase Ur, Richard Shay, Yang Wang, Rebecca Balebako, and Lorrie Cranor. "Why Johnny Can't Opt Out: A Usability Evaluation of Tools to Limit Online Behavioral Advertising." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '12. Austin, TX, USA: ACM, 2012, pp. 589–598. DOI: 10.1145/2207676.2207759. URL: https://dl.acm.org/doi/10.1145/2207676.2207759.

[155] Ada Lerner, Anna Kornfeld Simpson, Tadayoshi Kohno, and Franziska Roesner. "Internet Jones and the Raiders of the Lost Trackers: An Archaeological Study of Web Tracking from 1996 to 2016." In: *Proceedings of the 25th USENIX Security Symposium*. USENIX Security '16. Austin, TX, USA: USENIX Association, 2016, pp. 997–1013. URL: https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/lerner.

[156] Christophe Leung, Jingjing Ren, David Choffnes, and Christo Wilson. "Should You Use the App for That? Comparing the Privacy Implications of App- and Web-based Online Services." In: *Proceedings of the 2016 Internet Measurement Conference*. IMC '16. Santa Monica, CA, USA: ACM, 2016, pp. 365–372. DOI: 10.1145/2987443.2987456. URL: https://dl.acm.org/doi/10.1145/2987443.2987456.

[157] Frank Li, Zakir Durumeric, Jakub Czyz, Mohammad Karami, Michael Bailey, Damon McCoy, Stefan Savage, and Vern Paxson. "You've Got Vulnerability: Exploring Effective Vulnerability Notifications." In: *Proceedings of the 25th USENIX Security Symposium*. USENIX Security

'16. Austin, TX, USA: USENIX Association, 2016, pp. 1033–1050. URL: https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/li.

[158] Tianshi Li, Yuvraj Agarwal, and Jason I. Hong. "Coconut: An IDE Plugin for Developing Privacy-Friendly Apps." In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2.4 (Dec. 2018), pp. 1–35. DOI: 10.1145/3287056. URL: https://dl.acm.org/doi/10.1145/3287056.

[159] Tianshi Li, Elizabeth Louie, Laura Dabbish, and Jason I. Hong. "How Developers Talk About Personal Data and What It Means for User Privacy: A Case Study of a Developer Forum on Reddit." In: *Proceedings of the ACM on Human-Computer Interaction* 4.CSCW3 (Dec. 2020), pp. 1–28. DOI: 10.1145/3432919. URL: https://dl.acm.org/doi/10.1145/3432919.

[160] Timothy Libert. "Exposing the Hidden Web: An Analysis of Third-Party HTTP Requests on 1 Million Websites." In: *International Journal of Communication* 9 (2015), pp. 3544–3561. URL: https://ijoc.org/index.php/ijoc/article/view/3646.

[161] Timothy Libert. "An Automated Approach to Auditing Disclosure of Third-Party Data Collection in Website Privacy Policies." In: *The Web Conference 2018 – Proceedings of The World Wide Web Conference*. WWW '18. Lyon, France: IW3C2, 2018, pp. 207–216. DOI: 10.1145/3178876.3186087. URL: https://dl.acm.org/doi/10.1145/3178876.3186087.

[162] Timothy Libert and Rasmus Kleis Nielsen. *Third-Party Web Content on EU News Sites: Potential Challenges and Paths to Privacy Improvement*. May 2018. URL: https://timlibert.me/pdf/Libert_Nielsen-2018-Third_Party_Content_EU_News_GDPR.pdf.

[163] Thomas Linden, Rishabh Khandelwal, Hamza Harkous, and Kassem Fawaz. "The Privacy Policy Landscape After the GDPR." In: *Proceedings on Privacy Enhancing Technologies* 2020.1 (Jan. 2020), pp. 47–64. DOI: 10.2478/popets-2020-0004. URL: https://petsymposium.org/popets/2020/popets-2020-0004.php.

[164] Chao Liu, Ryen W. White, and Susan Dumais. "Understanding Web Browsing Behaviors Through Weibull Analysis of Dwell Time." In: *33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '10. Geneva, Switzerland: ACM, 2010, pp. 379–386. DOI: 10.1145/1835449.1835513. URL: https://dl.acm.org/doi/10.1145/1835449.1835513.

[165] Fei Liu, Rohan Ramanath, Norman Sadeh, and Noah A. Smith. "A Step Towards Usable Privacy Policy: Automatic Alignment of Privacy Statements." In: *Proceedings of the 25th International Conference on Computational Linguistics*. COLING 2014. Dublin, Ireland: Dublin City University and ACL, 2014, pp. 884–894. URL: https://aclanthology.org/C14-1084/.

[166] Kai-Uwe Loser and Martin Degeling. "Security and Privacy as Hygiene Factors of Developer Behavior in Small and Agile Teams." In: *ICT and Society – 11th IFIP TC 9 International Conference on Human Choice and Computers*. HCC 2014. Turku, Finland: Springer, 2014, pp. 255–265. DOI: 10.1007/978-3-662-44208-1_21. URL: https://link.springer.com/chapter/10.1007/978-3-662-44208-1_21.

[167] Max Maass, Marc-Pascal Clement, and Matthias Hollick. "Snail Mail Beats Email Any Day: On Effective Operator Security Notifications in the Internet." In: *Proceedings of the 16th International Conference on Availability, Reliability and Security*. ARES 2021. Vienna, Austria: ACM, 2021, pp. 1–13. DOI: 10.1145/3465481.3465743. URL: https://dl.acm.org/doi/10.1145/3465481.3465743.

[168] Max Maass, Henning Pridöhl, Dominik Herrmann, and Matthias Hollick. "Best Practices for Notification Studies for Security and Privacy Issues on the Internet." In: *Proceedings of the 16th International Conference on Availability, Reliability and Security*. ARES 2021. Vienna, Austria: ACM, 2021, pp. 1–10. DOI: 10.1145/3465481.3470081. URL: https://dl.acm.org/doi/10.1145/3465481.3470081.

[169] Max Maass, Alina Stöver, Henning Pridöhl, Sebastian Bretthauer, Dominik Herrmann, Matthias Hollick, and Indra Spiecker. "Effective Notification Campaigns on the Web: A Matter of Trust, Framing, and Support." In: *Proceedings of the 30th USENIX Security Symposium*. USENIX Security '21. Virtual Event, USA: USENIX Association, 2021, pp. 2489–2506. URL: https://www.usenix.org/conference/usenixsecurity21/presentation/maass.

[170] Manafactory. *Ginger – EU Cookie Law*. 2019. URL: https://wordpress.org/plugins/ginger/.

[171] Veronica Marotta, Vibhanshu Abhishek, and Alessandro Acquisti. "Online Tracking and Publishers' Revenues: An Empirical Analysis." In: *2019 Workshop on the Economics of Information Security*. WEIS '19. Boston, MA, USA, 2019. URL: https://weis2019.econinfosec.org/wp-content/uploads/sites/6/2019/05/WEIS_2019_paper_38.pdf.

[172] Kirsten Martin. "Do Privacy Notices Matter? Comparing the Impact of Violating Formal Privacy Notices and Informal Privacy Norms on Consumer Trust Online." In: *The Journal of Legal Studies* 45.S2 (June 2016), S191–S215. DOI: 10.1086/688488. URL: https://www.journals.uchicago.edu/doi/10.1086/688488.

[173] Arunesh Mathur, Gunes Acar, Michael Friedman, Elena Lucherini, Jonathan Mayer, and Marsh Chetty. "Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites." In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW (Nov. 2019), pp. 1–32. DOI: 10.1145/3359183. URL: https://dl.acm.org/doi/10.1145/3359183.

[174] Célestin Matte, Nataliia Bielova, and Cristiana Santos. "Do Cookie Banners Respect my Choice? Measuring Legal Compliance of Banners from IAB Europe's Transparency and Consent Framework." In: *Proceedings of the 2020 IEEE Symposium on Security and Privacy*. SP '20. San Francisco, CA, USA: IEEE Computer Society, 2020, pp. 791–809. DOI: 10.1109/SP40000.2020.00076. URL: https://ieeexplore.ieee.org/document/9152617.

[175] MaxMind, Inc. *GeoLite2 Free Geolocation Data*. 2022. URL: https://dev.maxmind.com/geoip/geolite2-free-geolocation-data.

[176] McAfee, LLC. *Customer URL Ticketing System*. 2022. URL: https://www.trustedsource.org/.

[177] David McCabe and Matina Stevis-Gridneff. *U.S. and European leaders reach deal on trans-Atlantic data privacy*. Mar. 2022. URL: https://www.nytimes.com/2022/03/25/business/us-europe-data-privacy.html.

[178] Peter McCullagh and John A. Nelder. *Generalized Linear Models*. 2nd ed. London, United Kingdom: Chapman & Hall, 1989. ISBN: 978-0-41-231760-6. URL: https://www.routledge.com/Generalized-Linear-Models/McCullagh-Nelder/p/book/9780412317606.

[179] Aleecia M. McDonald and Lorrie Faith Cranor. "The Cost of Reading Privacy Policies." In: *I/S: A Journal of Law and Policy for the Information Society* 4.3 (2008), pp. 543–568. URL: https://kb.osu.edu/handle/1811/72839.

[180] Aleecia M. McDonald and Lorrie Faith Cranor. "Americans' Attitudes About Internet Behavioral Advertising Practices." In: *9th Annual ACM Workshop on Privacy in the Electronic Society*. WPES '10. Chicago, IL, USA: ACM, 2010, pp. 63–72. DOI: 10.1145/1866919.1866929. URL: https://dl.acm.org/doi/10.1145/1866919.1866929.

[181] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. "Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice." In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW (Nov. 2019), pp. 1–23. DOI: 10.1145/3359174. URL: https://dl.acm.org/doi/10.1145/3359174.

[182] William Melicher, Mahmood Sharif, Joshua Tan, Lujo Bauer, Mihai Christodorescu, and Pedro Giovanni Leon. "(Do Not) Track Me Sometimes: Users' Contextual Preferences for Web Tracking." In: *Proceedings on Privacy Enhancing Technologies* 2016.2 (2016), pp. 135–154. DOI: 10.1515/popets-2016-0009. URL: https://petsymposium.org/popets/2016/popets-2016-0009.php.

[183] Abraham H. Mhaidli, Yixin Zou, and Florian Schaub. ""We Can't Live Without Them!" App Developers' Adoption of Ad Networks and Their Considerations of Consumer Risks." In: *Proceedings of the Fifteenth Symposium on Usable Privacy and Security*. SOUPS 2019. Santa Clara,

CA, USA: USENIX Association, 2019, pp. 225–244. URL: https://www.usenix.org/conference/soups2019/presentation/mhaidli.

[184] Vikas Mishra, Pierre Laperdrix, Antoine Vastel, Walter Rudametkin, Romain Rouvoy, and Martin Lopatka. "Don't Count Me Out: On the Relevance of IP Address in the Tracking Ecosystem." In: *The Web Conference 2020 – Proceedings of The World Wide Web Conference*. WWW '20. Taipei, Taiwan: IW3C2, 2020, pp. 808–815. DOI: 10.1145/3366423.3380161. URL: https://dl.acm.org/doi/abs/10.1145/3366423.3380161.

[185] Moove Agency. *GDPR Cookie Compliance*. Version 1.2.6. 2018. URL: https://wordpress.org/plugins/gdpr-cookie-compliance/.

[186] Alex Moss and Marco Milesi. *EU Cookie Law*. Version 3.0.5. 2018. URL: https://wordpress.org/plugins/eu-cookie-law/.

[187] Sarah Nadi, Stefan Krüger, Mira Mezini, and Eric Bodden. "'Jumping Through Hoops': Why do Java Developers Struggle With Cryptography APIs?" In: *Proceedings of the 38th International Conference on Software Engineering*. ICSE '16. Austin, TX, USA: ACM, 2016, pp. 935–946. DOI: 10.1145/2884781.2884790. URL: https://dl.acm.org/doi/10.1145/2884781.2884790.

[188] Alena Naiakshina, Anastasia Danilova, Eva Gerlitz, Emanuel von Zezschwitz, and Matthew Smith. "'If you want, I can store the encrypted password.' A Password-Storage Field Study with Freelance Developers." In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. Glasgow, United Kingdom: ACM, 2019. DOI: 10.1145/3290605.3300370. URL: https://dl.acm.org/citation.cfm?id=3300370.

[189] Alena Naiakshina, Anastasia Danilova, Christian Tiefenau, Marco Herzog, Sergej Dechand, and Matthew Smith. "Why Do Developers Get Password Storage Wrong? A Qualitative Usability Study." In: *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. CCS '17. Dallas, TX, USA: ACM, 2017, pp. 311–328. DOI: 10.1145/3133956.3134082. URL: https://dl.acm.org/citation.cfm?id=3134082.

[190] Alena Naiakshina, Anastasia Danilova, Christian Tiefenau, and Matthew Smith. "Deception Task Design in Developer Password Studies: Exploring a Student Sample." In: *Proceedings of the Fourteenth Symposium on Usable Privacy and Security*. SOUPS '18. Baltimore, MD, USA: USENIX Association, 2018, pp. 297–313. URL: https://www.usenix.org/conference/soups2018/presentation/naiakshina.

[191] National Institute of Standards and Technology. *NIST Privacy Framework*. 2020. URL: https://www.nist.gov/privacy-framework.

[192] John A. Nelder and Robert W. M. Wedderburn. "Generalized Linear Models." In: *Journal of the Royal Statistical Society, Series A* 135.3 (1972), pp. 370–384. DOI: 10.2307/2344614. URL: https://www.jstor.org/stable/2344614.

[193] Nick Nikiforakis, Luca Invernizzi, Alexandros Kapravelos, Steven Van Acker, Wouter Joosen, Christopher Kruegel, Frank Piessens, and Giovanni Vigna. "You Are What You Include: Large-scale Evaluation of Remote JavaScript Inclusions." In: *Proceedings of the 2012 ACM Conference on Computer and Communications Security*. CCS 2012. Raleigh, NC, USA: ACM, 2012, pp. 736–747. DOI: 10.1145/2382196.2382274. URL: https://dl.acm.org/doi/10.1145/2382196.2382274.

[194] Razieh Nokhbeh Zaeem and K. Suzanne Barber. "A study of web privacy policies across industries." In: *Journal of Information Privacy and Security* 13.4 (2017), pp. 169–185. DOI: 10.1080/15536548.2017.1394064. URL: https://www.tandfonline.com/doi/full/10.1080/15536548.2017.1394064.

[195] Midas Nouwens, Ilaria Liccardi, Michael Veale, David Karger, and Lalana Kagal. "Dark Patterns after the GDPR: Scraping Consent Pop-ups and Demonstrating their Influence." In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. Honolulu, HI, USA: ACM, 2020, pp. 1–13. DOI: 10.1145/3313831.3376321. URL: https://dl.acm.org/doi/10.1145/3313831.3376321.

[196] noyb. *News Sites: Readers need to "buy back" their own data at an exorbitant price*. Aug. 2021. URL: https://noyb.eu/en/news-sites-readers-need-buy-back-their-own-data-exorbitant-price.

[197] noyb. *noyb files 422 formal GDPR complaints on nerve-wrecking "Cookie Banners"*. Aug. 2021. URL: https://noyb.eu/en/noyb-files-422-formal-gdpr-complaints-nerve-wrecking-cookie-banners.

[198] noyb. *Austrian DSB: Use of Google Analytics violates "Schrems II" decision by CJEU*. Jan. 2022. URL: https://noyb.eu/en/austrian-dsb-eu-us-data-transfers-google-analytics-illegal.

[199] Sean O'Connor, Ryan Nurwono, Aden Siebel, and Eleanor Birrell. "(Un)clear and (In)conspicuous: The Right to Opt-out of Sale under CCPA." In: *Proceedings of the 20th Workshop on Workshop on Privacy in the Electronic Society*. WPES '21. Virtual Event, Republic of Korea: ACM, 2021, pp. 59–72. DOI: 10.1145/3463676.3485598. URL: https://dl.acm.org/doi/10.1145/3463676.3485598.

[200] Mike O'Neill. *Do Not Track and the GDPR*. June 2018. URL: https://www.w3.org/blog/2018/06/do-not-track-and-the-gdpr/.

[201] Lara O'Reilly. *Google's new CAPTCHA security login raises 'legitimate privacy concerns'*. Feb. 2015. URL: https://www.businessinsider.com.au/google-no-captcha-adtruth-privacy-research-2015-2.

[202] Jonathan A. Obar and Anne Oeldorf-Hirsch. "The biggest lie on the Internet: ignoring the privacy policies and terms of service policies of social networking services." In: *Information, Communication & Society* 23.1 (2020), pp. 128–147. DOI: 10.1080/1369118X.2018.1486870. URL: https://www.tandfonline.com/doi/full/10.1080/1369118X.2018.1486870.

[203] Objectis Ltd. *Cookie Script: free GDPR compliance for cookies*. 2018. URL: https://cookie-script.com/.

[204] OneTrust, LLC. *Cookie Consent and Website Scanning Products*. 2018. URL: https://www.onetrust.com/products/cookies/.

[205] heise online. *Shariff – Give Social Media Buttons Some Privacy*. 2019. URL: https://github.com/heiseonline/shariff.

[206] heise online. *embetty*. 2020. URL: https://github.com/heiseonline/embetty.

[207] PageFair. *The state of the blocked web – 2017 Global Adblock Report*. Tech. rep. Dublin, Ireland: PageFair Limited, Feb. 2017. URL: https://web.archive.org/web/20191130050941/https://pagefair.com/downloads/2017/01/PageFair-2017-Adblock-Report.pdf.

[208] Marcin Pajdzik. *EU Cookie Compliance*. Version 7.x-1.25. 2018. URL: https://www.drupal.org/project/eu_cookie_compliance.

[209] Mathias Panzenböck. *Social Share Privacy*. July 2019. URL: https://panzi.github.io/SocialSharePrivacy.

[210] Panagiotis Papadopoulos, Nicolas Kourtellis, and Evangelos Markatos. "Cookie Synchronization: Everything You Always Wanted to Know But Were Afraid to Ask." In: *The Web Conference 2019 – Proceedings of The World Wide Web Conference*. WWW '19. San Francisco, CA, USA: IW3C2, 2019, pp. 1432–1442. DOI: 10.1145/3308558.3313542. URL: https://dl.acm.org/doi/abs/10.1145/3308558.3313542.

[211] Panagiotis Papadopoulos, Peter Snyder, Dimitrios Athanasakis, and Benjamin Livshits. "Keeping out the Masses: Understanding the Popularity and Implications of Internet Paywalls." In: *The Web Conference 2020 – Proceedings of The World Wide Web Conference*. WWW '20. Taipei, Taiwan: IW3C2, 2020, pp. 1433–1444. DOI: 10.1145/3366423.3380217. URL: https://dl.acm.org/doi/10.1145/3366423.3380217.

[212] PayPal. *List of Third Parties (other than PayPal Customers) with Whom Personal Information May be Shared*. Oct. 2022. URL: https://www.paypal.com/ie/webapps/mpp/ua/third-parties-list.

[213] Mariana Peixoto, Dayse Ferreira, Mateus Cavalcanti, Carla Silva, Jéssyka Vilela, João Araújo, and Tony Gorschek. "On Understanding How Developers Perceive and Interpret Privacy Requirements Research Preview." In: *International Working Conference on Requirements Engineering: Foundation for Software Quality*. REFSQ 2020. Pisa, Italy: Springer Nature Switzerland AG, 2020, pp. 116–123. DOI: 10.1007/978-3-030-44429-7_. URL: https://link.springer.com/chapter/10.1007/978-3-030-44429-7_8.

[214] Pew Research Center. *October 2013 Higher Education and Gender Survey*. Oct. 2013. URL: https://www.pewsocialtrends.org/wp-content/uploads/sites/3/2014/02/higher-ed_topline.pdf.

[215] Aidan Polese, Safwat Hassan, and Yuan Tian. "Adoption of Third-party Libraries in Mobile Apps: A Case Study on Open-source Android Applications." In: *Proceedings of the 9th IEEE/ACM International Conference on Mobile Software Engineering and Systems 2022*. MOBILESoft 2022. Pittsburgh, PA, USA: ACM, 2022, pp. 125–135. DOI: 10.1145/3524613.3527810. URL: https://dl.acm.org/doi/10.1145/3524613.3527810.

[216] Barry Pollard. *Should you self-host Google Fonts?* Feb. 2020. URL: https://www.tunetheweb.com/blog/should-you-self-host-google-fonts/.

[217] Mathieu Pollet. *France joins Austria in finding Google Analytics illegal*. Feb. 2022. URL: https://www.euractiv.com/section/data-protection/news/france-joins-austria-says-google-analytics-data-not-protected-in-us/.

[218] Tara Poteat and Frank Li. "Who You Gonna Call? An Empirical Evaluation of Website security.txt Deployment." In: *Proceedings of the 21st ACM Internet Measurement Conference*. IMC '21. Virtual Event, USA: ACM, 2021, pp. 526–532. DOI: 10.1145/3487552.3487841. URL: https://dl.acm.org/doi/10.1145/3487552.3487841.

[219] Sören Preibusch, Thomas Peetz, Gunes Acar, and Bettina Behrendt. "Shopping for privacy: Purchase details leaked to PayPal." In: *Electronic Commerce Research and Applications* 15 (Jan. 2016), pp. 52–64. DOI: 10.1016/j.elerap.2015.11.004. URL: https://lirias.kuleuven.be/retrieve/385871.

[220] Quantcast. *GDPR Consent & Transparency Management*. 2018. URL: https://www.quantcast.com/gdpr/consent-management-solution/.

[221] Vladimir Radnaev. *GDPR Tools*. Version 1.0.2. 2018. URL: https://wordpress.org/plugins/gdpr-tools/.

[222] Ashwini Rao, Florian Schaub, Norman Sadeh, Alessandro Acquisti, and Ruogo Kang. "Expecting the Unexpected: Understanding Mismatched Privacy Expectations Online." In: *Proceedings of the Twelfth Symposium On Usable Privacy and Security*. SOUPS '16. Denver, CO, USA: USENIX Association, 2016, pp. 77–96. URL: https://www.usenix.org/conference/soups2016/technical-sessions/presentation/rao.

[223] Robert W. Reeder, Adrienne Porter Felt, Sunny Consolvo, Nathan Malkin, Christopher Thompson, and Serge Egelman. "An Experience Sampling Study of User Reactions to Browser Warnings in the Field." In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18. Montréal, QC, Canada: ACM, 2018, pp. 1–13. DOI: 10.1145/3173574.3174086. URL: https://dl.acm.org/citation.cfm?id=3174086.

[224] Joel R. Reidenberg et al. "Disagreeable Privacy Policies: Mismatches between Meaning and Users' Understanding." In: *Berkeley Technology Law Journal* 30.1 (2015), pp. 39–88. DOI: 10.2139/ssrn.2418297. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2418297.

[225] Harry Roberts. *Self-Host Your Static Assets*. May 2019. URL: https://csswizardry.com/2019/05/self-host-your-static-assets/.

[226] Franziska Roesner, Tadayoshi Kohno, and David Wetherall. "Detecting and Defending Against Third-Party Tracking on the Web." In: *Proceedings of the 9th USENIX Symposium on Networked Systems Design and Implementation*. NDSI '12. San Jose, CA, USA: USENIX Association, 2012. URL: https://dl.acm.org/doi/10.5555/2228298.2228315.

[227] Sebastian Roth, Timothy Barron, Stefano Calzavara, Nick Nikiforakis, and Ben Stock. "Complex Security Policy? A Longitudinal Analysis of Deployed Content Security Policies." In: *Proceedings of the 2020 Network and Distributed System Security Symposium*. NDSS '20. San Diego, CA, USA: Internet Society, 2020. DOI: 10.14722/ndss.2020.23046. URL: https://www.ndss-symposium.org/ndss-paper/complex-security-policy-a-longitudinal-analysis-of-deployed-content-security-policies/.

[228] Sebastian Roth, Lea Gröber, Michael Backes, Katharina Krombholz, and Ben Stock. "12 Angry Developers – A Qualitative Study on Developers' Struggles with CSP." In: *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. CCS '21. Virtual Event, Republic of Korea: ACM, 2021, pp. 3085–3103. DOI: 10.1145/3460120.3484780. URL: https://dl.acm.org/doi/abs/10.1145/3460120.3484780.

[229] Daniel Rücker and Tobias Kugler. *New European General Data Protection Regulation*. 1st ed. Baden-Baden, Germany: C. H. Beck / Hart / Nomos, 2018. ISBN: 978-3-848-73262-3. URL: https://beck-online.beck.de/Dokument?vpath=bibdata%2Fkomm%2FRueKugHdbEGDPR_1%2Fcont%2FRueKugHdbEGDPR.htm.

[230] Johnny Ryan. *Research result: what percentage will consent to tracking for advertising?* Sept. 2017. URL: https://web.archive.org/web/20181127023050/https://pagefair.com/blog/2017/new-research-how-many-consent-to-tracking/.

[231] Johnny Ryan. *French regulator shows deep flaws in IAB's consent framework and RTB*. Nov. 2018. URL: https://brave.com/cnil-consent-rtb/.

[232] Pasquale Salza, Fabio Palomba, Dario Di Nucci, Cosmo D'Uva, Andrea De Lucia, and Filomena Ferrucci. "Do Developers Update Third-Party Libraries in Mobile Apps?" In: *Proceedings of the 2018 IEEE/ACM 26th International Conference on Program Comprehension*. ICPC 2018. Gothenburg, Sweden: ACM, 2018, pp. 255–265. DOI: 10.1145/3196321.3196341. URL: https://ieeexplore.ieee.org/document/8972993.

[233] Pasquale Salza, Fabio Palomba, Dario Di Nucci, Andrea De Lucia, and Filomena Ferrucci. "Third-party libraries in mobile apps – When, how, and why developers update them." In: *Empirical Software Engineering* 25.3 (May 2020), pp. 2341–2377. DOI: 10.1007/s10664-019-09754-1. URL: https://link.springer.com/article/10.1007/s10664-019-09754-1.

[234] Iskander Sanchez-Rola, Matteo Dell'Amico, Platon Kotzias, Davide Balzarotti, Leyla Bilge, Pierre-Antoine Vervier, and Igor Santos. "Can I Opt Out Yet? GDPR and the Global Illusion of Cookie Control." In: *Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security*. AsiaCCS '19. Auckland, New Zealand: ACM, 2019, pp. 340–351. DOI: 10.1145/3321705.3329806. URL: https://dl.acm.org/doi/10.1145/3321705.3329806.

[235] Aaron Sankin and Surya Mattu. *The High Privacy Cost of a "Free" Website*. Sept. 2020. URL: https://themarkup.org/blacklight/2020/09/22/blacklight-tracking-advertisers-digital-privacy-sensitive-websites.

[236] Cristiana Santos, Nataliia Bielova, and Célestin Matte. "Are cookie banners indeed compliant with the law? Deciphering EU legal requirements on consent and technical means to verify compliance of cookie banners." In: *Technology and Regulation* 2 (Nov. 2020), pp. 91–135. DOI: 10.26116/techreg.2020.009. URL: https://techreg.org/article/view/10990.

[237] Florian Schaub, Rebecca Balebako, and Lorrie Faith Cranor. "Designing Effective Privacy Notices and Controls." In: *IEEE Internet Computing* 21.3 (May 2017), pp. 70–77. DOI: 10.1109/MIC.2017.75. URL: https://ieeexplore.ieee.org/document/7927931.

[238] Florian Schaub, Rebecca Balebako, Adam L. Durity, and Lorrie Faith Cranor. "A Design Space for Effective Privacy Notices." In: *Proceedings of the Eleventh Symposium On Usable Privacy and Security*. SOUPS '15. Ottawa, ON, Canada: USENIX Association, 2015, pp. 1–17. DOI: 10.1145/567752.567774. URL: https://www.usenix.org/conference/soups2015/proceedings/presentation/schaub.

[239] Quirin Scheitle, Oliver Hohlfeld, Julien Gamba, Jonas Jelten, Torsten Zimmermann, Stephen D. Strowes, and Narseo Vallina-Rodriguez. "A Long Way to the Top: Significance, Structure, and Stability of Internet Top Lists." In: *Proceedings of the Internet Measurement Conference 2018*. IMC '18. Boston, MA, USA: AMC, 2018, pp. 478–493. DOI: 10.1145/3278532.3278574. URL: https://dl.acm.org/doi/10.1145/3278532.3278574.

[240] Michel Schreiner and Thomas Hess. "Why Are Consumers Willing to Pay for Privacy? An Application of the Privacy-freemium Model to Media Companies." In: *Proceedings of the Twenty-Third European Conference on Information Systems*. ECIS '15. Münster, Germany: AIS, 2015. DOI: 10.18151/7217470. URL: http://aisel.aisnet.org/ecis2015_cr/164.

[241] Katharine Schwab. *Google's new reCAPTCHA has a dark side*. June 2019. URL: https://www.fastcompany.com/90369697/googles-new-recaptcha-has-a-dark-side.

[242] Ari Schwartz. *Looking Back at P3P: Lessons for the Future*. Nov. 2009. URL: https://cdt.org/wp-content/uploads/pdfs/P3P_Retro_Final_0.pdf.

[243] Adam Scott. *Building web apps that respect a user's privacy and security*. Jan. 2017. URL: https://www.oreilly.com/content/building-web-apps-that-respect-a-users-privacy-and-security/.

[244] Awanthika Senarath and Nalin A. G. Aarachchilage. "Understanding Software Developers' Approach towards Implementing Data Minimization." In: *4th Workshop on Security Information Workers*. WSIW 2018. Baltimore, MD, USA: USENIX Association, 2018. URL: https://wsiw2018.l3s.uni-hannover.de/papers/wsiw2018-Senarath.pdf.

[245] Awanthika Senarath and Nalin A. G. Arachchilage. "Why developers cannot embed privacy into software systems? An empirical investigation." In: *Proceedings of the 22nd International Conference on Evaluation and Assessment in Software Engineering 2018*. EASE '18. Christchurch, New Zealand: ACM, 2018, pp. 211–216. DOI: 10.1145/3210459.3210484. URL: https://dl.acm.org/doi/10.1145/3210459.3210484.

[246] Suranga Seneviratne, Harini Kolamunna, and Aruna Seneviratne. "A Measurement Study of Tracking in Paid Mobile Applications." In: *Proceedings of the 8th ACM Conference on Security & Privacy in Wireless and Mobile Networks*. WiSec '15. New York, NY, USA: ACM, 2015. DOI: 10.1145/2766498.2766523. URL: https://dl.acm.org/doi/abs/10.1145/2766498.2766523.

[247] Mario Silic. "Understanding Colour Impact on Warning Messages: Evidence from US and India." In: *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. CHI EA '16. San Jose, CA, USA: ACM, 2016, pp. 2954–2960. DOI: 10.1145/2851581.2892276. URL: https://dl.acm.org/citation.cfm?id=2892276.

[248] Silktide Ltd. *Cookie Consent by Insites – The most popular solution to the EU cookie law*. 2018. URL: https://cookieconsent.insites.com/.

[249] Michael Simon. *Apple is removing the Do Not Track toggle from Safari, but for a good reason*. Feb. 2019. URL: https://www.macworld.com/article/232426/apple-safari-removing-do-not-track.html.

[250] Software Freedom Conservancy. *Selenium Web Driver*. 2022. URL: https://www.selenium.dev/documentation/webdriver/.

[251] Jannick Sørensen and Sokol Kosta. "Before and After GDPR: The Changes in Third Party Presence at Public and Private European Websites." In: *The Web Conference 2019 – Proceedings of The World Wide Web Conference*. WWW '19. San Francisco, CA, USA: IW3C2, 2019, pp. 1590–1600. DOI: 10.1145/3308558.3313524. URL: https://dl.acm.org/doi/10.1145/3308558.3313524.

[252] Wissem Soussi, Maciej Korczyński, Sourena Maroofi, and Andrzej Duda. "Feasibility of Large-Scale Vulnerability Notifications after GDPR." In: *Proceedings of the 5th IEEE European Symposium on Security and Privacy Workshops*. EUROS&PW 2020. Virtual Event, Italy: IEEE Computer Society, 2020, pp. 531–536. DOI: 10.1109/EuroSPW51379.2020.00078. URL: https://ieeexplore.ieee.org/document/9229722.

[253] Katta Spiel, Oliver Haimson, and Danielle Lottridge. "How to do better with gender on surveys: A guide for HCI researchers." In: *Interactions* 26.4 (July 2019), pp. 62–65. DOI: 10.1145/3338283. URL: https://dl.acm.org/doi/10.1145/3338283.

[254] Stack Overflow. *2021 Developer Survey*. 2021. URL: https://insights.stackoverflow.com/survey/2021.

[255] Jake Stanich. *Simple Cookie Compliance*. Version 7.x-1.5. 2018. URL: https://www.drupal.org/project/simple_cookie_compliance.

[256] Oleksii Starov, Phillipa Gill, and Nick Nikiforakis. "Are You Sure You Want to Contact Us? Quantifying the Leakage of PII via Website Contact Forms." In: *Proceedings on Privacy Enhancing Technologies* 2016.1 (Jan. 2016), pp. 20–33. DOI: 10.1515/popets-2015-0028. URL: https://petsymposium.org/popets/2016/popets-2015-0028.php.

[257] State of California Department of Justice, Office of the Attorney General. *California Consumer Privacy Act (CCPA) – Frequently Asked Questions*. 2022. URL: https://oag.ca.gov/privacy/ccpa.

[258] State of California Legislative Counsel. *Assembly Bill No. 375 – Chapter 55*. June 2018. URL: https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180AB375.

[259] Marius Steffens, Marius Musch, Martin Johns, and Ben Stock. "Who's Hosting the Block Party? Studying Third-Party Blockage of CSP and SRI." In: *Proceedings of the 28th Annual Network and Distributed System Security Symposium*. NDSS '21. Virtual Event, USA: Internet Society, 2021. DOI: 10.14722/ndss.2021.24028. URL: https://www.ndss-symposium.org/ndss-paper/whos-hosting-the-block-party-studying-third-party-blockage-of-csp-and-sri/.

[260] Ben Stock, Giancarlo Pellegrino, Frank Li, Michael Backes, and Christian Rossow. "Didn't You Hear Me? – Towards More Successful Web Vulnerability Notifications." In: *25th Annual Network and Distributed System Security Symposium*. NDSS '18. San Diego, CA, USA: Internet Society, 2018. URL: https://www.ndss-symposium.org/wp-

content/uploads/2018/02/ndss2018_01B-1_Stock_paper.pdf.

[261] Ben Stock, Giancarlo Pellegrino, Christian Rossow, Martin Johns, and Michael Backes. "Hey, You Have a Problem: On the Feasibility of Large-Scale Web Vulnerability Notification." In: *Proceedings of the 25th USENIX Security Symposium*. USENIX Security '16. Austin, TX, USA: USENIX Association, 2016, pp. 1015–1032. URL: https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/stock.

[262] Samuel Stolton. *GDPR enforcement held back by lack of resources, report says*. May 2020. URL: https://www.euractiv.com/section/data-protection/news/gdpr-enforcement-held-back-by-lack-of-resources-report-says/.

[263] Tim Strack. *Use of Google Analytics without anonymizeIP is a violation of data protection law*. Jan. 2020. URL: https://web.archive.org/web/20210420080029/https://www.lhr-law.de/en/magazine/use-of-google-analytics-without-anonymizeip-is-a-violation-of-data-protection-law/.

[264] Christian Stransky, Yasemin Acar, Duc Cuong Nguyen, Dominik Wermke, Elissa M. Redmiles, Doowon Kim, Michael Backes, Simson Garfinkel, Michelle L. Mazurek, and Sascha Fahl. "Lessons Learned from Using an Online Platform to Conduct Large-Scale, Online Controlled Security Experiments with Software Developers." In: *10th USENIX Workshop on Cyber Security Experimentation and Test*. CSET '17. Vancouver, BC, Canada: USENIX Association, 2017. URL: https://www.usenix.org/conference/cset17/workshop-program/presentation/stransky.

[265] Mohammad Tahaei, Alisa Frik, and Kami Vaniea. "Deciding on Personalized Ads: Nudging Developers About User Privacy." In: *Proceedings of the Seventeenth Symposium on Usable Privacy and Security*. SOUPS 2021. Virtual Event, USA: USENIX Association, 2021, pp. 573–595. URL: https://www.usenix.org/system/files/soups2021-tahaei.pdf.

[266] Mohammad Tahaei, Kopo M. Ramokapane, Tianshi Li, Jason I. Hong, and Awais Rashid. "Charting App Developers' Journey Through Privacy Regulation Features in Ad Networks." In: *Proceedings on Privacy Enhancing Technologies* 2022.3 (2022), pp. 33–56. DOI: 10.56553/popets-2022-0061. URL: https://petsymposium.org/popets/2022/popets-2022-0061.php.

[267] Mohammad Tahaei and Kami Vaniea. ""Developers Are Responsible": What Ad Networks Tell Developers About Privacy." In: *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI EA '21. Virtual Event, Japan: ACM, 2021, pp. 1–11. DOI: 10.1145/3411763.3451805. URL: https://dl.acm.org/doi/fullHtml/10.1145/3411763.3451805.

[268] Mohammad Tahaei, Kami Vaniea, Konstantin Beznosov, and Maria K. Wolters. "Security Notifications in Static Analysis Tools: Developers' Attitudes, Comprehension, and Ability to Act on Them." In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI '21. Virtual Event, Japan: ACM, 2021, pp. 1–17. DOI: 10.1145/3411764.3445616. URL: https://dl.acm.org/doi/10.1145/3411764.3445616.

[269] Mohammad Tahaei, Kami Vaniea, and Naomi Saphra. "Understanding Privacy-Related Questions on Stack Overflow." In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. Honolulu, HI, USA: ACM, 2020, pp. 1–14. DOI: 10.1145/3313831.3376768. URL: https://dl.acm.org/doi/10.1145/3313831.3376768.

[270] TeleTrusT and ENISA. *IT Security Act (Germany) and EU General Data Protection Regulation: Guideline "State of the art" – Technical and organisational measures*. Tech. rep. V 1.9_2021-09 EN. Berlin, Germany: IT Security Association Germany (TeleTrusT), Sept. 2021. URL: https://www.teletrust.de/fileadmin/user_upload/2021-09_TeleTrusT_Guideline_State_of_the_art_in_IT_security_EN.pdf.

[271] Welderufael B. Tesfay, Peter Hofmann, Toru Nakamura, Shinsaku Kiyomoto, and Jetzabel Serna. "PrivacyGuide: Towards an Implementation of the EU GDPR on Internet Privacy Policy Evaluation." In: *Proceedings of the Fourth ACM International Workshop on Security and Privacy Analytics*. IWSPA '18. Tempe, AZ, USA: ACM, 2018, pp. 15–21. DOI: 10.1145/3180445.3180447. URL: https://dl.acm.org/doi/10.1145/3180445.3180447.

[272] Richard H. Thaler and Cass R. Sunstein. *Nudge: Improving Decisions About Health, Wealth, and Happiness*. 1st ed. New Haven, CT, USA: Yale University Press, 2008. ISBN: 978-0-14-311526-7. URL: https://yalebooks.yale.edu/book/9780300122237/.

[273] The Tor Project. *Research Safety Board*. 2022. URL: https://research.torproject.org/safetyboard/.

[274] Wiebke Thode, Joachim Griesbaum, and Thomas Mandl. "'I would have never allowed it': User Perception of Third-party Tracking and Implications for Display Advertising." In: *Proceedings of the 14th International Symposium on Information Science*. ISI 2015. Zadar, Croatia: Verlag Werner Hülsbusch, 2015, pp. 445–456. DOI: 10.5281/zenodo.17971. URL: https://zenodo.org/record/17971.

[275] Christian Tiefenau, Emanuel von Zezschwitz, Maximilian Häring, Katharina Krombholz, and Matthew Smith. "A Usability Evaluation of Let's Encrypt and Certbot: Usable Security Done Right." In: *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. CCS '19. London, United Kingdom: ACM, 2019, pp. 1971–1988. DOI: 10.1145/3319535.3363220. URL: https://dl.acm.org/doi/10.1145/3319535.3363220.

[276] Emanuele Toscano. *cookieBAR – A free and easy cookie law plugin*. 2018. URL: https://cookie-bar.eu/.

[277] TrustArc Inc. *TrustArc Cookie Consent Manager*. 2018. URL: https://www.trustarc.com/products/consent-manager/.

[278] Joseph Turow, Michael Hennessy, and Nora Draper. "Persistent Misperceptions: Americans' Misplaced Confidence in Privacy Policies, 2003–2015." In: *Journal of Broadcasting & Electronic Media* 62.3 (July 2018), pp. 461–478. DOI: 10.1080/08838151.2018.1451867. URL: https://www.tandfonline.com/doi/full/10.1080/08838151.2018.1451867.

[279] Tobias Urban, Martin Degeling, Thorsten Holz, and Norbert Pohlmann. "'Your Hashed IP Address: Ubuntu.' Perspectives on Transparency Tools for Online Advertising." In: *Proceedings of the 35th Annual Computer Security Applications Conference*. ACSAC '19. San Juan, Puerto Rico, USA: ACM, 2019, pp. 702–717. DOI: 10.1145/3359789.3359798. URL: https://dl.acm.org/doi/10.1145/3359789.3359798.

[280] Tobias Urban, Martin Degeling, Thorsten Holz, and Norbert Pohlmann. "Beyond the Front Page: Measuring Third Party Dynamics in the Field." In: *The Web Conference 2020 – Proceedings of The World Wide Conference*. WWW '20. Taipei, Taiwan: IW3C2, 2020, pp. 1275–1286. DOI: 10.1145/3366423.3380203. URL: https://dl.acm.org/doi/abs/10.1145/3366423.3380203.

[281] Tobias Urban, Dennis Tatang, Martin Degeling, Thorsten Holz, and Norbert Pohlmann. "A Study on Subject Data Access in Online Advertising After the GDPR." In: *Proceedings of the 14th DPM International Workshop on Data Privacy Management*. DPM '19. Luxembourg, Luxembourg: Springer Nature Switzerland AG, 2019, pp. 61–79. DOI: 10.1007/978-3-030-31500-9_5. URL: https://link.springer.com/chapter/10.1007/978-3-030-31500-9_5.

[282] Tobias Urban, Dennis Tatang, Martin Degeling, Thorsten Holz, and Norbert Pohlmann. "Measuring the Impact of the GDPR on Data Sharing in Ad Networks." In: *Proceedings of the 15th ACM Asia Conference on Computer and Communications Security*. ASIA CCS '20. Taipei, Taiwan: ACM, 2020, pp. 222–235. DOI: 10.1145/3320269.3372194. URL: https://dl.acm.org/doi/10.1145/3320269.3372194.

[283] Pelayo Vallina, Victor Le Pochat, Álvaro Feal, Marius Paraschiv, Julien Gamba, Tim Burke, Oliver Hohlfeld, Juan Tapiador, and Narseo Vallina-Rodriguez. "Mis-shapes, Mistakes, Misfits: An Analysis of Domain Classification Services." In: *Proceedings of the 2020 ACM Internet Measurement Conference*. IMC '20. Virtual Event, USA: ACM, 2020, pp. 598–618. DOI: 10.1145/3419394.3423660. URL: https://dl.acm.org/doi/10.1145/3419394.3423660.

[284] Rob van Eijk, Hadi Asghari, Philipp Winter, and Arvind Narayanan. "The Impact of User Location on Cookie Notices (Inside and Outside of the European Union)." In: *Workshop on Technology and Consumer Protection*. ConPro '19. San Francisco, CA, USA: IEEE Computer So-

ciety, 2019. URL: https://www.ieee-security.org/TC/SPW2019/ConPro/papers/vaneijk-conpro19.pdf.

[285] Maggie Van Nortwick and Christo Wilson. "Setting the Bar Low: Are Websites Complying With the Minimum Requirements of the CCPA?" In: *Proceedings on Privacy Enhancing Technologies* 2022.1 (Jan. 2022), pp. 608–628. DOI: 10.2478/popets-2022-0030. URL: https://petsymposium.org/popets/2022/popets-2022-0030.php.

[286] Marie Vasek and Tyler Moore. "Do Malware Reports Expedite Cleanup? An Experimental Study." In: *5th Workshop on Cyber Security Experimentation and Test*. CSET '12. Bellevue, WA, USA: USENIX Association, 2012. URL: https://www.usenix.org/conference/cset12/workshop-program/presentation/Vasek.

[287] Thomas Vissers, Wouter Joosen, and Nick Nikiforakis. "Parking Sensors: Analyzing and Detecting Parked Domains." In: *Proceedings of the 2015 Network and Distributed System Security Symposium*. NDSS 2015. San Diego, CA, USA: Internet Society, 2015. DOI: 10.14722/ndss.2015.23053. URL: https://www.ndss-symposium.org/ndss2015/ndss-2015-programme/parking-sensors-analyzing-and-detecting-parked-domains/.

[288] Tim Wambach and Katharina Bräunlich. "The Evolution of Third-Party Web Tracking." In: *Proceedings of the Second International Conference on Information Systems Security and Privacy*. ICISSP 2016. Rome, Italy: Springer Nature Switzerland AG, 2016, pp. 130–147. DOI: 10.1007/978-3-319-54433-5_8. URL: https://link.springer.com/chapter/10.1007/978-3-319-54433-5_8.

[289] The Web Incubator Community Group. *Web Monetization*. May 2022. URL: https://webmonetization.org/.

[290] WebToffee. *GDPR Cookie Consent Plugin*. Version 1.7.1. 2018. URL: https://www.webtoffee.com/product/gdpr-cookie-consent/.

[291] Markus Weinmann, Christoph Schneider, and Jan vom Brocke. "Digital Nudging." In: *Business & Information Systems Engineering* 58.6 (Dec. 2016), pp. 433–436. DOI: 10.1007/s12599-016-0453-1. URL: https://link.springer.com/article/10.1007/s12599-016-0453-1.

[292] Ben Weinshel, Miranda Wei, Mainack Mondal, Euirim Choi, Shawn Shan, Claire Dolin, Michelle L. Mazurek, and Blase Ur. "Oh, the Places You've Been! User Reactions to Longitudinal Transparency About Third-Party Web Tracking and Interferencing." In: *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. CCS '19. London, United Kingdom: ACM, 2019, pp. 149–166. DOI: 10.1145/3319535.3363200. URL: https://dl.acm.org/doi/10.1145/3319535.3363200.

[293] WidgetPack. *Cookie Law Bar*. Version Version 1.2.1. 2018. URL: https://wordpress.org/plugins/cookie-law-bar/.

[294] Chamila Wijayarathna and Nalin A. G. Arachchilage. "Why Johnny Can't Store Passwords Securely? A Usability Evaluation of Bouncycastle Password Hashing." In: *Proceedings of the 22nd International Conference on Evaluation and Assessment in Software Engineering 2018*. EASE '18. Christchurch, New Zealand: ACM, 2018, pp. 205–210. DOI: 10.1145/3210459.3210483. URL: https://dl.acm.org/doi/10.1145/3210459.3210483.

[295] Shomir Wilson et al. "The Creation and Analysis of a Website Privacy Policy Corpus." In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. ACL '16. Berlin, Germany: ACL, 2016, pp. 1330–1340. DOI: 10.18653/v1/P16-1126. URL: https://aclanthology.org/P16-1126/.

[296] Carl Woodhouse. *jquery.cookieBar*. 2018. URL: https://carlwoodhouse.com/jquery.cookieBar/.

[297] World Economic Forum. *Redesigning Data Privacy: Reimagining Notice & Consent for human-technology interaction*. July 2020. URL: https://www3.weforum.org/docs/WEF_Redesigning_Data_Privacy_Report_2020.pdf.

[298] wunderfarm. *WF Cookie Consent*. Version 1.1.4. 2018. URL: https://wordpress.org/plugins/wf-cookie-consent/.

[299] Eric Zeng, Frank Li, Emily Stark, Adrienne Porter Felt, and Parisa Tabriz. "Fixing HTTPS Misconfigurations at Scale: An Experiment with Security Notifications." In: *The 2019 Workshop on the Economics of Information Security*. WEIS '19. Boston, MA, USA, 2019. URL: https://weis2019.econinfosec.org/wp-content/uploads/sites/6/2019/05/WEIS_2019_paper_16.pdf.

[300] Zippia. *Web Developer Demographics and Statistics in the US*. Apr. 2022. URL: https://www.zippia.com/web-developer-jobs/demographics/.