Saarland University
Faculty of Mathematics and Computer Science
Department of Computer Science

Reproducible and Responsible Web Security and Privacy Measurement Research

Dissertation
zur Erlangung des Grades
des Doktors der Ingenieurwissenschaften
der Fakultät für Mathematik und Informatik
der Universität des Saarlandes

von
Florian Hantke

Saarbrücken,
2026

Tag des Kolloquiums:	26.06.2026
Dekan:	Prof. Dr. Jan Reineke

Prüfungsausschuss:

Vorsitzender:	Prof. Dr. Martina Maggio
Berichterstattende:	Prof. Dr.-Ing. Ben Stock
	Prof. Dr.-Ing. Christoph Sorge
	Prof. Adam Doupé
Akademischer Mitarbeiter:	Dr. Sylvain Chatel

Zusammenfassung

Reproduzierbarkeit und Verantwortung von Forschung sind Grundlage guter Wissenschaft, stellen jedoch im Bereich der Websicherheit oft Hürden dar. In dieser Dissertation beleuchten wir beide Aspekte im Kontext von Web Measurements.

Im ersten Teil konzentrieren wir uns auf Reproduzierbarkeit von Forschung und untersuchen, inwiefern Webarchive als Datenquellen für Web Security & Privacy Measurements geeignet sind. Wir vergleichen Datenqualität und zeigen grundlegende Probleme auf, etwa mangelnde Genauigkeit und Defizite mit dynamischen Inhalten. Zur Überwindung dieser Probleme stellen wir ein neues Vorgehen vor, das Seitenaktivitäten erfasst und reproduzierbarere Web Measurements ermöglicht.

Im zweiten Teil widmen wir uns der Forschungsverantwortung. Wir betrachten rechtliche und ethische Überlegungen bei Web Measurements und diskutieren Perspektiven von verschiedenen Expert:innen. Darauf aufbauend skizzieren wir Handlungsstrategien für Forschende und demonstrieren diese in einer Fallstudie über Access Control Issues. Abschließend werfen wir einen Blick auf führende Konferenzen und ihre Ziele und Praktiken hinsichtlich der Förderung verantwortungsvoller Forschung.

Zusammenfassend bietet diese Dissertation Strategien für Forschende, um Web Measurements durchzuführen, die sowohl reproduzierbar als auch verantwortungsvoll sind.

Abstract

Reproducibility and responsibility are the foundation of rigorous scientific research, yet both present challenges for web security and privacy researchers. In this dissertation, we view these challenges through the lens of large-scale web measurements.

We first focus on reproducibility, exploring how web archives can serve as valuable data sources for web security and privacy measurements. We analyze which archives provide the most comprehensive data for researchers, while also highlighting obstacles these archives present, such as limited accuracy and problems with dynamic content. To address these shortcomings, we present a new tool and format that captures page activities, enabling more reproducible measurements.

We then turn to responsibility, examining the legal and ethical considerations involved in measuring web security and privacy at scale. Drawing on insights from an interview study with legal experts, ethics committee members, and site operators, we outline strategies one can use when facing these challenges. We demonstrate the application of these strategies in a case study on access control issues. Lastly, we take a look at the side of top-tier conferences, to see what their goals and practices are in terms of fostering responsible research ethics.

Together, these contributions provide strategies for researchers to conduct web security and privacy measurements that are both reproducible and responsible.

UNIVERSITÄT DES SAARLANDES

Eidesstattliche Versicherung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet. Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form in einem Verfahren zur Erlangung eines akademischen Grades vorgelegt.

Declaration of original authorship

I hereby declare that this dissertation is my own original work except where otherwise indicated. All data or concepts drawn directly or indirectly from other sources have been correctly acknowledged. This dissertation has not been submitted in its present or similar form to any other academic institution either in Germany or abroad for the award of any other degree.

Saarbrücken, 03/2026

gez. / signed

Florian Hantke

Background of this Dissertation

This dissertation is based on the papers listed below, for all of which I was the lead author ([P1, P2, P3, P4]). Throughout this work, I benefited from close and stimulating collaborations with colleagues from multiple institutions. All papers were published or are under submission at leading peer-reviewed conferences in security and privacy.

[P1] Hantke, F., Calzavara, S., Wilhelm, M., Rabitti, A., and Stock, B. You Call This Archaeology? Evaluating Web Archives for Reproducible Web Security Measurements. In: *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 2023. DOI: 10.1145/3576915.3616688.

[P2] Hantke, F., Roth, S., Mrowczynski, R., Utz, C., and Stock, B. Where Are the Red Lines? Towards Ethical Server-Side Scans in Security and Privacy Research. In: *2024 IEEE Symposium on Security and Privacy (SP)*. 2024. DOI: 10.1109/SP54263.2024.00104.

[P3] Hantke, F., Snyder, P., Haddadi, H., and Stock, B. Web Execution Bundles: Reproducible, Accurate, and Archivable Web Measurements. In: *34th USENIX Security Symposium (USENIX Security 25)*. 2025.

[P4] Hantke, F., Mrowczynski, R., and Stock, B. Reflection, Education, Consistency: Towards Best Ethics Practices At Security And Privacy Conferences. In: *Proceedings of the 2026 ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 2026.

Further Contributions of the Author

In addition to these primary works, I also contributed as a co-author to two further papers ([S1, S2]) and served as a supervisor on another ([S3]). At the beginning of my doctoral research, I published my Master's thesis at a leading measurement conference as first author [S4].

[S1] El Hajj Chehade, S., Hantke, F., and Stock, B. 403 Forbidden? Ethically Evaluating Broken Access Control in the Wild. In: *2025 IEEE Symposium on Security and Privacy (SP)*. 2025. DOI: 10.1109/SP61157.2025.00252.

[S2] Mustafa, A., Rautenstrauch, J., Hantke, F., Agarwal, S., Calzavara, S., and Stock, B. LeakyLinks: Measuring the Security and Privacy Risks of URL Scanning Services. In: *2026 IEEE Symposium on Security and Privacy (SP)*. 2026.

[S3] Decker, P. and Hantke, F. VDPCollect: Vulnerability Disclosure Programs as a Complement to Web Security Measurements. In: *Proceedings of the 21st ACM Asia Conference on Computer and Communications Security (AsiaCCS)*. 2026.

[S4] Hantke, F. and Stock, B. HTML Violations and Where to Find Them: A Longitudinal Analysis of Specification Violations in HTML. In: *Proceedings of the 22nd ACM Internet Measurement Conference (IMC)*. 2022. DOI: 10.1145/3517745.3561437.

Acknowledgments

Es ist nicht gut, daß der Mensch alleine sei, und besonders nicht, daß er alleine arbeite; vielmehr bedarf er der Teilnahme und Anregung, wenn etwas gelingen soll.

"It is not good for humans to be alone, and especially not to work alone; rather, he needs participation and stimulation if something is to succeed."

— Johann Wolfgang von Goethe

Following Goethe's advice, this thesis is anything but the product of solitary work. It is the result of years filled with inspiring discussions, generous support, and the love of fantastic people. Throughout this journey, I had the opportunity to attend many conferences, workshops, and retreats at which I always enjoyed lively discussions about all sorts of research. At home, I was lucky to get all the support I needed.

First and foremost, I thank my wife, Charly! You encouraged me to begin this journey of pursuing a PhD in the first place. Your endless support, especially during moments when challenges felt unsolvable and the workload exhausting, has been invaluable to me. Without you, these lines would not have been written.

I am deeply grateful to my family, Regina, Fritz, and Kilian, for preparing me for all challenges in life, for supporting me with helpful advice, and for never questioning the decisions I made in life. You always believed in me, and I know how proud you feel now. I also thank my parents-in-law, Elisabeth and Peter, for their thoughtful advice and many enriching conversations, ranging from international cuisines and journeys to the German legal system.

Of course, my sincere thanks go to Ben Stock for being the best supervisor I could have wished for. Through my role as doctoral representative, I have had the opportunity to see a wide range of supervision styles and approaches, and have thus far not seen anyone doing better. You always had my back, provided steady support when challenges emerged, and helped find solutions – even at midnight before my birthday.

But Ben would likely not be where he is today without Felix Freiling, whom I first had the pleasure of meeting as a lecturer during my Bachelor's studies. Your inspiring lectures have sparked my interest in research and ultimately set me on the path toward an academic career. You have remained a trusted and supportive mentor, always approachable and generous with advice, even throughout my PhD. Thank you, Felix!

I am also deeply grateful to my collaborators for their insightful discussions. In particular, I thank Rafael for introducing me to qualitative research methods, for discussing sociology and ethics, and for guiding me through the very first interviews I conducted. I am also thankful to Pete and Hamed for our collaborative project and my time as an intern at Brave. That experience offered me a valuable perspective on industry research and startup culture, opened my world to browser internals and privacy research, and significantly helped me to find out what path I want to take after

this dissertation. I also thank Stefano for his advice not only on research but also on reviewing papers. I further thank Bailey for his thoughtful advice and engaging chats on research ethics. The final paper in this thesis would have taken a very different form without your tips and advice in the beginning. And of course, my sincere thanks go to all my co-authors (in alphabetical order): Ali, Alvise, Ben, Christine, Jannis, Moritz, Philip, Shubham, Saiid, Sebastian, Simone, Stefano, Til.

My thanks also go to the SWAG group and everyone who was part of it during the time – mainly Sebastian, Jannis, Shubham, Tin, Christine, Ali, Kristina, and Valentino. Thank you for all the fruitful discussions and brainstorming sessions.

To my friends – especially Christoph, Max, Anna, and Niels – thank you for all the cheerful discussions about legal and political topics, geeking out about nerdy computer deep dives, and silly jokes. To Caro, Keno, and Lukas, thank you for the great evenings we had cooking whenever I was in Saarbrücken. To Jannis, thank you for the many fruitful chats about the Web, science, and future career plans, whenever we shared an apartment, from the very beginning, when we both began writing our Master's thesis.

I am grateful to the examiners – Ben, Christoph, and Adam – for taking the time to read and review this thesis and for their feedback, and thoughtful and curious questions.

Finally, I thank everyone at CISPA, from research colleagues to HR and PR, for creating an environment in which this journey became not only possible but truly enjoyable.

And to anyone who reads this while writing their own dissertation: this has been the most enjoyable and emotional part of the entire thesis – apart from the fear that I might have unintentionally forgotten someone!

Contents

List of Figures

List of Tables

Acronyms

3S Server-Side Scanning. 91

AC Access Control. 117

CSP Content Security Policy. 4

DOM Document Object Model. 13

ERB Ethical Review Board. 24

GDPR General Data Protection Regulation. 15

HTML Hypertext Markup Language. 12

HTTP Hypertext Transfer Protocol. 11

IA Internet Archive. 33

ICT Information and Communication Technology. 25

IRB Institutional Review Board. 24

JS JavaScript. 13

REC Research Ethics Committee. 24

S&P Security and Privacy. 3

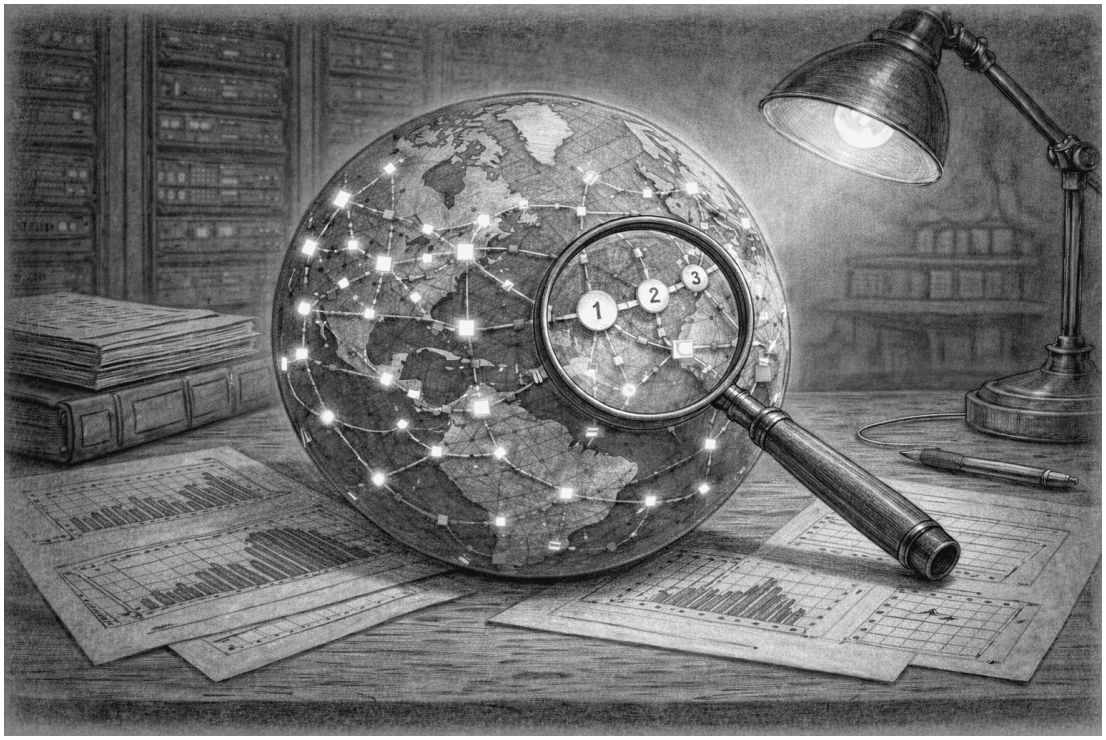
TLS Transport Layer Security. 4

VSF Variable Swapping Framework. 122

WebREC Web Execution Bundle RECorder. 68

1

Introduction



1.1 Reproducibility and Responsibility Challenges in Web Security and Privacy Measurement Research

Public confidence in science has shown signs of decline in recent years. In the United States, survey data indicate a decline from earlier levels of trust, particularly following the periods of high public confidence and attention during the Covid-19 pandemic [142]. While the confidence in science is still stable in Germany, a slight decrease in trust can also be observed there, most notably among individuals with an intermediate level of formal education [132]. Populism is being cited as one of the main contributing factors to this decline [265], yet studies further point to people's concerns about scientific views that may not align with their own beliefs, and also to ethical concerns – more broadly, questions regarding the *responsibility* of scientists [142]. For example, one survey shows that only 65% of the surveyed population perceives scientists as acting ethically, placing this attribute in the middle range compared to other professional characteristics [142]. In addition, concerns about human error, methodological bias, and insufficient transparency, and thus about *reproducibility* of scientific practices, further undermine confidence in research results [142, 192]. Although scientists cannot directly influence the broader political dynamics, such as populism, they can strengthen the credibility of science itself. This requires them to particularly understand more and care about the responsibility and the reproducibility of their scientific methods.

The questions around responsibility and reproducibility in scientific research are not new. Historical incidents like the Tuskegee Syphilis Study [30], or the human studies during World War II, have demonstrated how scientific progress detached from ethical responsibility can lead to severe harm. During World War II, Nazi physicians forced prisoners to undergo inhumane procedures and experiments in the name of clinical research, like extreme freezing or injections of typhus bacteria and other toxic substances [244, 63]. These abuses caused profound moral concerns about the role of science in society and prompted a response in the form of the Nuremberg Doctors' Trials, which in turn led to an increased awareness of ethical issues in research and efforts to improve oversight, such as the creation of Institutional Review Boards (IRBs).

In a similar vein, a more recent challenge concerns the reproducibility and replicability of scientific findings – the so-called reproducibility crisis, also known as replication crisis [105, 19]. Researchers have found that many published scientific results cannot be reproduced when other teams repeat the same studies, leading to findings of limited value. Widespread media coverage (e.g.,[263, 241]) suggests that this crisis may negatively affect public trust in science [202]. At the same time, the crisis has led to efforts to make science more transparent and reproducible, for example, by encouraging authors to pre-register studies and provide complete access to their study material [202]. Such open science practices have been shown to increase public trust in science [195]. Overall, an ongoing shift in culture within science towards valuing reproducibility more highly is recognizable [19].

These mentioned issues, often associated with social sciences, are also highly relevant in the domain of Security and Privacy (S&P) research, a highly applied field embedded in socio-technical systems involving users and organizations. Incidents such as the Hypocrite Commit paper [95], in which researchers attempted to submit malformed

commits to the Linux kernel, as well as voices criticizing the state of reproducibility in the S&P research domain [205, 71, 126, 57], illustrate these concerns. In response, an increasing number of conferences (e.g., [18, 187, 139]) now require ethical considerations in papers, sections about open-science, and artifact evaluation. Moreover, workshops about meta research (e.g., [8]) are created, demonstrating an increasing commitment to self-reflection and continuous improvement within the S&P research domain to address reproducibility and responsibility challenges.

Reproducibility and responsibility challenges, of course, differ across research fields within the S&P domain. In this dissertation, we focus on web S&P research, specifically on measurement studies – empirical analysis of trends such as the adoption of Transport Layer Security (TLS) or the deployment and effectiveness of the Content Security Policy (CSP). The Web is a platform used by most individuals daily for activities ranging from travel planning and social communication to interactions with public administrations or banks. As an integral part of everyday life, measuring and understanding evolving issues with S&P trends on the Web is critical to improve how society is protected against malicious activities (e.g., through informed campaigns that promote TLS or CSP adoption). At the same time, this ubiquity makes web S&P measurement research ethically challenging as flaws in methods quickly affect hundreds of thousands of users. Moreover, the fast evolution of websites and technologies complicates making findings reproducible; yesterday's findings may already look different today (e.g., due to the browsers' shift from warning to blocking or upgrading mixed content as we learn later). These challenges motivate a closer look at current practices.

In this dissertation, we address these two central challenges, reproducibility and responsibility in web security and privacy measurement research, to improve the public confidence in our research. With respect to reproducibility, we establish an evidence-based understanding of the problem and propose and evaluate solutions in the form of web archives and a novel archiving format. On the responsibility side, we look into the challenges regarding ethics and law in server-side measurement research, broaden the understanding of how to conduct such type of research in a responsible way, and lastly analyze what top-tier S&P conferences currently do – and could do – to foster ethical considerations among S&P researchers. Overall, our goal is to improve the way web S&P measurement researchers plan and execute their studies to enable them to deliver impactful results while meeting the highest standard of reproducibility, ethical responsibility, and thus improve the public confidence in their findings.

1.2 Contributions

This thesis is motivated by the question of *how to conduct reproducible and responsible web security and privacy measurements*. To answer this question, we split the question into two threads. First, we discuss the problems of reproducibility, propose and evaluate web archives as a solution by analyzing which is the best source for archival data and whether we can trust their correctness (RQ1). Following observations of issues with dynamic web content on archives, we then study how an accurate, reproducible, and general-purpose archiving format for web security and privacy measurements should look and thus propose a new web archiving tool and format and evaluate its accuracy

improvements compared to traditional archiving formats (RQ2).

With respect to responsibility, we first discuss ethical and legal challenges in server-side web measurement research and answer the question of how this kind of research could be conducted responsibly and lawfully (RQ3). Lastly, we zoom out and ask the broader question of what our communities' perceived goals are regarding ethics-fostering interventions and how top-tier conferences do in regard to promoting ethical considerations and ethics-fostering interventions (RQ4).

RQ1: What Are the Best Sources of Archival Data Available to Date and Can We Trust Their Correctness?

To address the reproducibility challenge in web security and privacy measurement studies, we systematically investigate the potential of using web archives as a reproducible data source. We first evaluate 38 web archives by querying 5,000 highly ranked domains, demonstrating how the Internet Archive clearly outperforms its competitors in terms of freshness and completeness. Next, we assess how correct the data on the Internet Archive is for historical web security measurements, since reproducibility is only meaningful if the underlying data is accurate. This evaluation highlights several limitations and potential pitfalls that researchers need to keep in mind. Finally, we investigate whether recent archival data of the Internet Archive could serve as a substitute for live security measurements. Our results indicate that archive-based security and privacy measurements make a promising, reproducible-by-design alternative to traditional live measurements. At the same time, they show inherent challenges, such as data originating from various uncontrolled vantage points with unknown configurations, or dynamically negotiated web content. Based on these findings, we therefore identify insights and best practices for future archive-based measurement studies.

RQ2: How Should an Accurate, Reproducible, and General-Purpose Archiving Format for Web S&P Measurements Look Like?

To address the limitations identified in RQ1, we introduce an open-source web measurement tool, WebREC, together with a novel archiving format *.web*, that extends common archival data with additional page activities. We first present a comprehensive overview of web measurement and archive tools commonly used in research, along with their limitations. Based on this overview, we then compare WebREC against existing tools and show that it is more *accurate*, as it correctly measures and attributes events that cannot be captured with existing tools; more *general*, as its archived data is reusable for a broad range of measurements without modifications; and *comprehensive*, as it handles events from all relevant page behaviors. We empirically validate these claims by replicating past web measurement studies and experiments on tracking requests, JavaScript API calls, and DOM event handlers. The results support the accuracy improvements, especially the coverage of JavaScript APIs. The results show that *.web* captures nearly 100% of the analyzed JavaScript API calls, whereas HAR and WARC archives miss appearances for around 40% of the API calls. To assess the practical utility of WebREC and *.web* for the security and privacy community, we then analyze in the last part a set of papers presented in a 2024 web crawling SoK paper.

We find that 70% of the surveyed studies could be conducted using WebREC as is, and that 48% could directly use *.web* archives without requiring any new crawling.

RQ3: How Can We Enable Server-Side Scanning Research Within a Framework That Prevents Harm for All Stakeholders?

On the responsibility side, we first need to better understand challenges, potential solutions, and legal and ethical boundaries of web S&P measurements. To this end, we turn to challenging server-side scanning research and present a qualitative interview study with 23 participants, including legal experts, members of Research Ethics Committees (RECs), and website and server operators. The study explores diverse perspectives through a set of semi-structured questions and five typical server-side scanning scenarios. We complement the findings with a quantitative survey study of website operators, drawing from 119 responses. Our analysis across these stakeholder groups indicates that the lack of judicial precedent and the absence of clear ethical guidelines create substantial obstacles in overcoming the risks associated with server-side scanning. To support the community in addressing these challenges, we then present best practices for future server-side scanning research and demonstrate their applicability through a practical server-side scanning study that adheres to these practices.

RQ4: What Ethics-Fostering Interventions Do S&P Conferences Have and Need to Achieve Their Ethical Goals?

As RQ3 reveals, the research community lacks clear and widely adopted ethics guidelines. We therefore examine how top-tier conferences currently address ethics in practice and in their reviewing processes. To foster a shared understanding within our community and to inform conference-level decisions, we perform a content analysis of 99 calls for papers of the top four conferences and conduct in-depth, semi-structured interviews with 20 senior and junior community members, including 10 (former) chairs of programs and RECs. In these interviews, we explore reasons for ethics procedures, their intended goals, and discuss participants' experiences with existing ethics interventions, as well as potential new approaches proposed by our team. Building on these insights, we then present four identified goal groups, a collection of existing and proposed ethics-fostering interventions, and evaluate them against the identified goals and findings from the interviews. To bring the community forward, we argue for coordinated, community-wide ethics steering efforts and take a first step by introducing an open ethics wiki.

1.3 Open Science

To promote transparency and reproducibility in security and privacy research, we publish the code and artifacts used throughout this dissertation whenever possible. Some materials, such as interview transcripts, cannot be made publicly available in order to protect the privacy of interview participants.

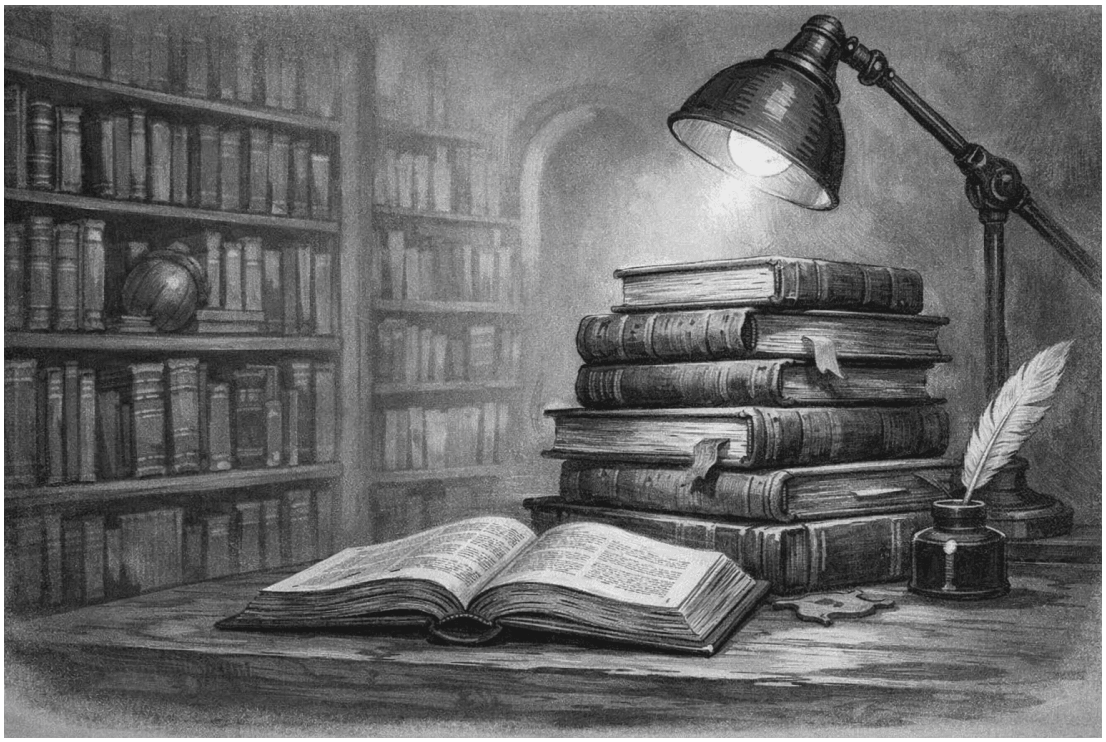
For Chapter 3, we publish our data collection code [87], for Chapter 4 we publish the pipeline of our experiments [91] and parts of the datasets to Zenodo [90]; full

datasets that are too large to host are available upon request. To support future web measurement research, we also maintain and publicly release our measurement tools from Chapter 4, including WebREC and its associated crawler and query framework [32, 33]. For the qualitative studies (Chapters 5 and 7), we publish supplementary materials such as interview guides, survey instruments, and codebooks used for analysis in the appendix (Appendix C and Appendix D) and on GitHub [92, 88].

1.4 Outline

This thesis is organized into eight chapters. Following this introduction, in Chapter 2, we provide the theoretical and technical background on web technology and measurements, and introduce the concepts related to reproducibility, and responsibility. In addition to that, we discuss prior work relevant to this thesis. Chapter 3 investigates web archives as a data source for reproducible web security and privacy measurements. It evaluates multiple archives, assesses the correctness of archival data for historical measurements, and analyzes whether recent archival data can serve as a substitute for live measurements. Following an observation of issues with dynamic data on these archives, Chapter 4 introduces the new archiving tool WebREC and format *.web*. It describes the system design and evaluates its accuracy in comparison to existing archival formats and measurement tools. Chapter 5 then moves on to discuss to responsibility in server-side web measurement research. It presents findings from interviews with legal experts, REC members, and web operators and concludes with best practices for conducting such measurements responsibly. Chapter 6 demonstrates the practical applicability of these best practices and insights through a server-side measurement of access control issues. Chapter 7 analyzes how top-tier conferences address ethics in practice. It presents a content analysis of call for papers, and semi-structured interviews with community members to explore reasons and goals of ethics procedures and to discuss existing and proposed ethics-fostering interventions. Finally, Chapter 8 concludes this thesis by reflecting on what we learned about reproducibility and responsibility, and elaborates on future work.

2
Background and Related Work



In this chapter, we present the background for this thesis and review the relevant literature. The chapter is organized into four parts. First, we introduce relevant web technologies and concepts of web security and privacy. Second, we describe what web measurements are and the methods used to measure them. Third, we examine reproducibility in the context of scientific research and specific challenges in web-based studies. Lastly, we define what we mean by responsibility in the context of this thesis and outline the ethical and legal background that guides our work.

Contribution of the Author

This chapter compiles background material from all the main papers [P1, P2, P3, P4, S1] presented in this thesis. These sections were originally contributed by the thesis author and various co-authors, and have been combined, extended, and adapted for this overview by the thesis author.

2.1 Web Security and Privacy

To understand web security and privacy concepts, we first give an overview of the core web technologies. Afterwards, we introduce client-side and server-side attacks, security headers for mitigating them, and concepts of web privacy.

2.1.1 The Web

The Web began as a decentralized network for sharing information among scientists. Since then, it has developed into a highly dynamic ecosystem which touches nearly every aspect of modern life [44]. Still, the core concept is simple: the Web consists of a vast number of web servers that provide resources and information, which can be requested and viewed by a web client such as a browser.

2.1.1.1 Hypertext Transfer Protocol

Communication between web servers and web clients is based on the Hypertext Transfer Protocol (HTTP), a protocol introduced in 1991 by Tim Berners-Lee and first standardized by the Internet Engineering Task Force in 1996 [25]. Via HTTP, a client sends a request to a server and receives a response back. A request, as in Listing 1, contains a method, a path, a set of headers, and optionally a message body. The method specifies the desired action, for example, retrieving a resource (`GET`), submitting data (`POST`), updating a resource (`PUT`), or deleting it (`DELETE`). The path identifies the resource on the server that the client requests or wants to interact with. Headers provide additional metadata, such as the addressed host (`Host`), the content length and format of transmitted data (`Content-Length`, `Content-Type`), or response formats expected by the client (`Accept`). If present, the message body, following an empty line after the headers, contains additional data transmitted to the server. In response, the server returns, as in Listing 2, a status code such as `200` for success, response headers similar to the request headers, and an optional message body.

Listing 1 Example for a HTTP Request.

```
1 GET /welcome.html HTTP/1.1
2 Host: example.com
3 Accept: text/html
4
```

Listing 2 Example for a HTTP Response.

```
1 HTTP/1.1 200 OK
2 Content-Type: text/html
3 Content-Length: 1337
4
5 <!doctype html>
6 [...]
```

Early HTTP traffic was transmitted in plain text, meaning that anyone with access to the communication channel could read or manipulate the exchanged data. To address this limitation, the Secure Sockets Layer (SSL) was introduced in 1994 and later evolved into Transport Layer Security (TLS). TLS provides encryption, authenticity, and integrity protection for HTTP traffic, commonly known as HTTP Secure (HTTPS).

Another aspect of HTTP is that it is a stateless protocol. Each request is processed independently, and the server does not store information about previous interactions. To enable stateful interactions such as authentication, modern web applications rely on additional mechanisms. One common approach is the use of cookies: small pieces of user-related information are stored on the client side and transmitted via the `Cookie` header with each request to the corresponding server. Other mechanisms include authentication headers stored in the browser's `localStorage`.

2.1.1.2 Hypertext Markup Language

To present the requested information to a web client in a structured format, the most common document format is Hypertext Markup Language (HTML). Modern documents follow the HTML Living Standard [254], maintained by the Web Hypertext Application Technology Working Group. The standard defines the syntax, semantics, and processing of HTML.

A typical HTML document (see Listing 3) begins with a document type declaration (`<!DOCTYPE html>`), followed by a hierarchical tree of HTML elements. Each element consists of an opening tag and, in most cases, a corresponding closing tag such as the HTML root element (`<html>`). Some elements are self-closing elements (e.g., `<img>`, `<br>`), meaning they do not require a closing tag. Many elements support attributes that further specify their behavior or reference external resources. For example, media elements, such as image and video elements (`<img>` and `<video>`), use the `src` attribute to indicate the location of the resource to be loaded.

Listing 3 Example for an HTML document.

```
1  <!DOCTYPE html>
2  <html lang="en">
3    <head>
4      <meta charset="utf-8" />
5      <title>Welcome Page</title>
6      <link rel="stylesheet" href="/styles/main.css" />
7      <script src="/js/main.js"></script>
8    </head>
9
10   <body>
11     <h1 id="header">Hello, Reader!</h1>
12     <img
13       src="/imgs/welcome.png"
14       onload="console.log('Img loaded!')">
15     <script>
16       console.log('Document loaded!');
17     </script>
18   </body>
19 </html>
```

2.1.1.3 JavaScript

To enable dynamic behavior and interactive functionality on a web page, developers use JavaScript (JS). Typically, JS code is either directly embedded inline within the HTML document using `<script>` elements (Listing 3 in line 15) or loaded from an external file referenced by a `src` attribute of a `<script>` element (Listing 3 in line 7). If external files are loaded from another domain, they are called *third-party* scripts. Another way to execute JS is to trigger it via (user) events with event handlers (e.g., `onclick` or `onload`). As Listing 3 and Listing 4 demonstrate, such handlers may be specified directly as element attributes or added programmatically via JS.

JS enables developers to inspect and modify the structure and content of a loaded HTML document. This is performed through the browser's internal, in-memory tree representation of the document, known as the Document Object Model (DOM) [162]. By accessing and modifying DOM nodes, scripts can dynamically update text, insert or remove elements, or change styling at runtime.

Beyond DOM manipulation, JS provides access to a rich collection of web APIs that give access to browser and device functionalities. For instance, developers can use `fetch` to send network requests, or use the Battery API to retrieve information about the device's battery state. These APIs enable web applications to implement complex behavior within the browser environment.

2.1.1.4 Browser Boundaries

Because web applications can get complex and often handle sensitive data, such as user session tokens stored in cookies, the Web and browsers require isolation mechanisms

Listing 4 Example for a JavaScript file.

```
let t = 'GlzIHRoYXQgb2YgY3VyaW9zaXR5fQ==';
alert(atob('SGFudGtlQ1RGe215IGNyaW1lI' + t));

const b=document.createElement("button");
b.textContent="Show Battery";
header.after(b);

b.onclick=()=>navigator.getBattery()
.then(x=>header.insertAdjacentHTML("afterend",
  `.<p>Battery: ${Math.round(x.level*100)}%</p>`
));
```

to prevent one application from accessing another application's data. The fundamental browser-enforced security mechanism for this purpose is the Same Origin Policy (SOP) [164], which establishes this security boundary at the level of an *origin*.

An origin is defined as a combination of a protocol, a hostname, and a port. Two resources are considered to have the same origin only if all three components match. Under SOP, data owned by an origin is isolated from read and write accesses by scripts running in a different origin, e.g., a script running at `https://foo.com/bar` cannot access the DOM or any storage of a page at `https://baz.com`.

When a web page loads external JS code using a `<script>` element, the loaded script executes with the origin of the rendered page, rather than with the origin from which the script was fetched. Consequently, such scripts inherit the full privileges of the including page and can access all resources permitted to that origin. This design enables the use of external libraries and content delivery via third-party servers, but it also poses a significant security risk by allowing the inclusion of untrusted scripts.

A related security concern arises when a secure web page (served over `https`) includes resources from an insecure web page (served over `http`). Such inclusions are referred to as *mixed content* [163]. Because insecure resources are delivered without encryption, they can potentially be modified by network attackers and thereby attack the security and integrity of a secure page.

Browser handling of mixed content has evolved over time. Prior to around mid-2010s, browsers would load insecure JS resources referenced via `http` on a secure web page [257, 239]. After this change, such insecure scripts were blocked and no longer executed on a secure page. This change strengthened the security of the Web by preventing insecure sub-resources from attacking the integrity of secure origins, yet made the reproducibility of older experiments challenging.

2.1.2 Web Privacy

With the Web becoming a platform for many of our everyday tasks, like shopping or banking, the user leaves traces of personal information that reveal aspects of their behavior, preferences, and identity. Various actors on the Web collect and analyze such data, often with the goal of enabling user profiling, targeted advertising and other

forms of behavioral tracking. Privacy can be understood as the ability of individuals to control how their personal information is collected, processed, and shared [21].

One central privacy risk on the Web is tracking. While there are various forms of tracking on the Web, in this thesis, we focus only on third-party tracking, a technique in which activity data is transmitted to external tracking domains. Such data transmission is typically performed via HTTP requests initiated by loaded (third-party) resources, such as scripts, images, or frames. With enough data collected across many sites, these tracking providers can create a detailed profile of the user.

To counter these tracking practices, a variety of privacy-enhancing technologies exist, such as ad and tracker blockers (e.g., uBlock Origin [84]) or privacy-focused web browsers (e.g., Brave [31]). These technologies reduce unwanted data collection by blocking requests to known tracking domains, thereby preventing sites from embedding such trackers. Request blocking is often based on community-collected filter lists such as the EasyList [66] or diconnect.me [59].

Privacy also plays a critical role in terms of researchers' responsibilities. Beyond technical concerns, ethical and legal aspects must be taken into account. Researchers often collect, analyze, and store web measurement data and must therefore consider potential impacts and adhere to data protection regulations, such as the General Data Protection Regulation (GDPR).

2.1.3 Client-Side Web Vulnerabilities

The large volume of private and confidential data that websites handle makes them attractive targets not only for tracking companies but also for malicious actors. The objective of an attacker is usually to find flaws in the implementation of a website and then either get access to confidential information, break the integrity of information, or threat their availability (CIA Triad). In this thesis, we do not aim to provide an exhaustive overview of client-side web vulnerabilities. Instead, we provide a brief summary of the vulnerabilities mentioned in this thesis and motivate the relevance of the security mechanisms discussed later.

2.1.3.1 Cross-Site Scripting (XSS)

Probably the most well-known web vulnerability is Cross-Site Scripting (XSS), a vulnerability that has been extensively studied in academic research [174, 235, 38, 118, 231, 122, 233, 22]. In an XSS attack, the attacker injects malicious script code into a web page, causing it to be executed in the context of the victim's browser and with the privileges of the vulnerable page's origin. Such injections may be caused by flaws in *server-side* input handling or from insecure *client-side* logic. They can be *reflected* per request or *stored* persistently on the server or client, enabling attackers to steal sensitive data, manipulate page content, or perform actions on behalf of the victim.

2.1.3.2 Clickjacking

Clickjacking is a vulnerability in which a victim is fooled into interacting with a hidden element of another web page [178, 40]. Typically, the attacker embeds the target page

within a transparent frame positioned above a benign-looking element the victim is fooled into clicking. For example, an attacker-controlled page may promise free in-game credits, if a button is clicked, while a transparent overlay causes the click to be delivered to an embedded YouTube page, resulting in the victim to subscribe to the attacker's channel. Although the attacker cannot directly access the victim's YouTube data due to SOP, clickjacking still enables unauthorized actions to be performed on behalf of the user.

2.1.3.3 DOM Clobbering

DOM clobbering is a vulnerability in which an attacker manipulates the structure of the DOM to overwrite JS variables that the web application uses [180, 116]. By injecting specially crafted HTML elements through unsanitized user input, an attacker can create DOM nodes with names and identifiers that collide with existing JS variables, thereby manipulating the code's execution (e.g., `<input name="isAdmin" value="true">` affects `let is_admin = window.isAdmin || false;`). This can lead to security bypasses and enable XSS-like exploitation, even if traditional XSS is not possible.

2.1.3.4 Cross-Site leaks (XS-Leaks)

Cross-Site leaks (XS-Leaks) are a class of vulnerabilities that enable adversaries to leak cross-origin information about a user, thereby threatening the user's privacy [179, 127, 190]. The vulnerability exploits side channels such as differences in response codes or frame counting. In a typical scenario, an attacker runs JS code on their controlled page and observes these side channels while making cross-site requests to learn information about the visiting victim. In the past, attacks demonstrated that attackers could determine whether a victim is logged in to a targeted private GitHub account or whether the Google Mail application contains emails with a certain subject.

2.1.4 Client-Side Security Mechanisms

To prevent or mitigate vulnerabilities such as those mentioned above, web browsers implement many security mechanisms. In addition to the earlier-mentioned SOP and mixed-content policy, HTTP security headers play a prominent role in web application security. These headers are sent in HTTP responses, setting the additional configurable security policies that the web browser can enforce. In this thesis, we consider five security headers, which are among the most popular in the wild. They also received significant attention from the research community [197, 131, 264, 72].

2.1.4.1 X-Frame-Options (XFO)

The XFO header allows a web page to restrict the set of pages authorized to load it within an iframe. This is useful to prevent clickjacking or other types of frame-based attacks [98, 40]. The XFO header (`X-Frame-Options`) can be set to three different values: (*i*) `SAMEORIGIN` only allows framing on web pages sharing the same origin of the page setting the header; (*ii*) `DENY` entirely forbids framing on every web page; (*iii*)

`ALLOW-FROM` *ori* only allows framing on web pages hosted on the origin *ori*. The last option is now deprecated and unsupported by modern web browsers.

2.1.4.2 Content Security Policy (CSP)

The CSP header allows a web page to enforce declarative security policies, addressing a range of different threats [227, 197]. In particular, existing literature identifies three key use cases for the CSP header (`Content-Security-Policy`): (*i*) *XSS mitigation*: CSP allows the specification of a set of allowed origins for remote script inclusions and can block a page's ability to run inline scripts, inline event handlers, and string-to-code transformation functions like eval, which are the most common XSS vectors; (*ii*) *Framing control*: the frame-ancestors directive of CSP is intended as a modern replacement of XFO since it allows the specification of a set of origins which are allowed to frame the page; (*iii*) *TLS enforcement*: CSP supports the `upgrade-insecure-requests` and `block-all-mixed-content` directives to forbid plain HTTP requests and enforce the use of HTTPS.

2.1.4.3 HTTP Strict Transport Security (HSTS)

As an alternative to CSP TLS enforcement, the HSTS header can be used to enforce the HTTPS protocol and upgrade all requests made via HTTP automatically to HTTPS [94, 131]. HSTS uses the `max-age` directive to specify the duration of protection in seconds, i.e., how long the web browser should perform the HTTP-to-HTTPS upgrade. The protection can be extended to all subdomains of the activating host using the `includeSubDomains` option. For example, if the homepage of `https://foo.com` activates HSTS with `max-age` equal to 31536000 and the `includeSubDomains` option, any request towards a subdomain of `foo.com` will be automatically upgraded to HTTPS for one year (31536000 seconds). To prevent attacks against the very first response before observing the HSTS header, site operators can ask for inclusion in the HSTS preload list[1], which major web browsers use to determine which communications they should automatically upgrade.

2.1.4.4 Cross-Origin Policies

Attacks like XS-Leaks frequently rely on the attacker having a handle to the victim application's window object [127, 190]. This can be done, e.g., by using the `window.open` functionality in browsers to open another domain and using the return value for that call to observe the other domain's behavior through side channels and timing. To protect a site from this, operators can nowadays specify the Cross-Origin-Opener-Policy (COOP) header [157]. This enables fine-grained control of whether the opener can retain a handle. It can be `unsafe-none`, i.e., any opener can keep the handle. Alternatively, it can be set to `same-origin`, `same-origin-allow-popus`, or `noopener-allow-popus`, allowing the opener to retain the handle if it shares the opened document's origin or other strict conditions.

[1] `https://hstspreload.org/`

In addition to COOP, sites can also ensure inclusion behavior for other types of content, e.g., images. To do so, they can set the Cross-Origin-Embedder-Policy (COEP) header [156] to either `require-corp` (requiring cross-origin resources to set the CORP header), `credentialless` (ensuring that cross-origin resources are requested without credentials, i.e., cookies) or `unsafe-none` (default behavior, it allows any inclusion).

The Cross-Origin-Resource-Policy (CORP) [158], in turn, allows an operator to disallow access from other origins unless requests are CORS-enabled. CORP can be set to `cross-origin` (allowing any page to include the resource), `same-origin` (only same-origin pages can embed the file), or `same-site`.

2.1.4.5 Referrer-Policy

Browsers send along the so-called `Referer` (sic) header in outgoing requests. This header indicates the URL of the page that caused the request, e.g., when clicking a link, it includes the page from which the user came. Similarly, when including subresources such as images or scripts, browsers also provide the current document's URL. This has obvious privacy implications, in particular in cases where session management is implemented through tokens in the URL. To overcome this privacy threat, browsers have long since implemented the more privacy-friendly `Origin` header [159]. Rather than the full URL, this header only contains the document's origin. With Referrer-Policy (RP) [161], a web operator can control if the Referer header is sent along and what information it shall contain. The option `unsafe-url` sends the full URL in all outgoing requests. In contrast, `origin` only provides the origin, whereas `no-referrer` strips the header entirely. Additional policies regulate cross-origin requests and protocol downgrades: `no-referrer-when-downgrade` sends the full URL if the resource's protocol is at least as secure the current URL (e.g., HTTPS to HTTPS); `origin-when-cross-origin` sends the full URL for same-origin requests and only the origin for cross-origin requests; `same-origin` sends the referrer only for same-origin requests; `strict-origin` only sends the origin if the protocol remains the same (and nothing in case of protocol downgrade); and `strict-origin-when-cross-origin` sends the full URL to same-origin resources, the origin to equal-protocol cross-origin URLS, and nothing in less secure protocols.

2.1.4.6 Permissions-Policy

Permissions-Policy (PP) [160] allows a website a fine-grained control over the APIs accessible by JS, both for first- and, more importantly, third-party resources. This allows site operators to constrain the use of sensitive features, such as geolocation or camera use. For example, by configuring the PP header as `geolocation=(self "trusted.com")`, the site's own origin resources, as well as those in frames coming from trusted.com, are allowed to access the device's geographical location. Note that the policy applies only to documents, i.e., if the first party includes a script from `untrusted.com`, it runs in the context of the first party. That is, it has the same capabilities as the first party (in this case, that script can access the geo location).

2.1.5 Server-Side Web Attacks

This chapter has so far focused mainly on client-side technologies and security issues. While crucial to the web ecosystem, a significant part of a web application's logic runs and stores sensitive data on the server side. Although web security researchers have been analyzing client-side vulnerabilities in the wild, studies often avoid analysis of server-side issues on a large scale due to legal concerns and uncertainties [76].

Server-side vulnerabilities are flaws in the design and implementation of a web application's underlying backend infrastructure. Such flaws can be exploited by attackers to compromise the earlier-mentioned CIA Triad not only for an individual user, but potentially for all users of an affected application. For example, XSS, widely studied on the client side [174, 235, 231, 38], can also be considered a server-side vulnerability if the vulnerable logic to inject the attacker-controlled data lies in the server backend. In the following, we present common classes of server-side vulnerabilities.

2.1.5.1 SQL Injections (SQLi)

SQL Injections (SQLi) [119] is a class of server-side vulnerabilities that occurs when user input is concatenated to database queries without proper validation or sanitization. In a vulnerable case, attackers can craft payloads that escape the intended SQL statement and inject arbitrary SQL commands. Successful SQL injection attacks can allow attackers to read sensitive data from a database, manipulate information, or, in the worst case, execute operating system commands to gain full control over the server.

2.1.5.2 Path Traversal

Path traversal [175] is a server-side vulnerability in which attackers can control input data used to construct file system paths without proper validation or sanitization. It allows for manipulating path parameters (e.g., by using sequences such as `../`) to escape the intended directory context and access arbitrary files on the server. Attackers can use this attack to exfiltrate the application's source code, access sensitive files, or read other operating system information.

2.1.5.3 Authentication, Authorization, and Access Control Issues

In the context of web applications, *authentication* refers to verifying a user's identity, whereas *authorization* checks whether a user is permitted to perform a specific interaction or access data. Flaws in the implementation of either mechanism can allow unprivileged individuals to access information or perform actions they are not allowed to do, so-called *access control* (*AC*) issues.

One common instance of this vulnerability class is the Insecure Direct Object Reference (IDOR) [176]. IDOR vulnerabilities occur when a web application exposes references to data objects, for example, via an ID without validating user authorization (e.g., user 1337 can access `/profile/403/messages` without permission). Attackers can exploit this by simply enumerating such IDs and thereby accessing or modifying data belonging to other users.

Broken access control vulnerabilities are widespread in practice and regularly appear among the most critical web security risks [177]. Numerous incidents demonstrate their impact; for instance, a flaw in Instagram that allowed attackers to access phone numbers and user details by abusing a contact import API endpoint [93], or a vulnerability in NordVPN that allowed attackers to change user IDs and leak users' payment histories [54].

2.2 Web Measurements

Web measurements are a central methodology for understanding the prevalence of security and privacy flaws and mitigation strategies. Examples of prior work span from demonstrating the prevalence and impact of critical vulnerabilities like XSS [136, 235], studying mitigation techniques like HTTP Strict Transport Security (HSTS) [131] and cookie security attributes [35], criticizing fundamental design flaws in policies like CSP [250, 197], up to studying the adoption of privacy-related issues and regulations like the web tracking ecosystem [67] and cookie consent banners [146]. Such measurements can be used as empirical evidence in political discussions [210, 213] or the standardization process of web technologies [253].

2.2.1 Web Measurement Techniques

Researchers use a variety of web measurement techniques. These techniques range from sending simple HTTP requests with command-line tools like `curl` to large-scale crawling infrastructures that orchestrate instrumented browsers like Puppeteer [83]. Observed artifacts typically include HTTP traffic, headers, or JavaScript usage. Each technique has its own trade-offs, which we discuss in more detail in Chapter 4.2. Despite their differences, all techniques have one thing in common: they collect and observe data from a large number of websites (usually using lists such as Tranco [135] or CrUX [200]) to highlight the relevance of the findings.

In the context of this thesis, we distinguish between two broad classes of measurements. *Client-side scans* collect and analyze client-side artifacts (e.g., security headers and client-side JS code executions) that are observable within the browser during normal page execution or under controlled client-side modifications. On the other hand, we define *server-side scans* (3S) as an automated series of specially crafted requests and probes to a large number of web servers and sites to systematically assess them for specific server-side behavior, such as potential vulnerabilities like SQL injections or similar vulnerabilities.

2.2.2 Web Archives

Some measurements, in particular longitudinal experiments that aim to understand behavior from the past and over time (e.g., remote JavaScript inclusion [170], or CSP [197]), leverage *Web archives*. Web archives are services that periodically store copies of public web pages, associate them with their archival date, and make them available to requesting clients. A variety of archives exist on the Web; however, in this thesis, we primarily focus on those supporting the standard Memento archiving

protocol [222, 221]. Our focus on Memento has the advantage of defining a systematic criterion for identifying existing web archives for use in Chapter 3, with the additional benefit that they can be accessed using the same standardized protocol.

In addition, we consider Common Crawl as an alternative archival source. Common Crawl collects large portions of the Web monthly and stores them as snapshots. At the time of our study, the October 2022 snapshot contained over 2.55 billion pages, with an index exceeding 2TB, and has been used in prior web security measurements [197, S4]. The archived content is stored in compressed files containing WARC records with metadata (e.g., request time, content type, and size) followed by the captured content.

In the area of web archives, there exists academic work; however, it is largely based on the Internet Archive (IA) alone, without any strong motivation, and provides just limited insights on how to use it correctly for security analyses. The studies presented in this dissertation instead consider a plethora of different web archives, compare their effectiveness and accuracy, and present potential pitfalls of archive-based security measurements. In a sense, our analysis extends to the web security setting prior work on web privacy by Lerner et al., leveraging the IA to study the evolution of web tracking from 1996 to 2016 [138]. Their work has already documented challenges in using the IA, such as the bubble escapes that return responses from the live web. However, we extend their work by evaluating an extensive list of web archives and not just the IA, identifying new pitfalls in the use of the IA coming from the existence of different contributors therein, and investigating the use of the IA as a reproducible alternative to traditional live security measurements.

Further studies also examined the quality and quantity of archival data. Murphy et al. presented the first validity study of the IA with respect to archived web pages, website age, and website updates in 2008 [165]. More recent studies took a more in-depth look into the accuracy of archived copies of popular web pages, detecting that only one out of five pages actually existed as presented [5], and proposed solutions to mitigate this problem [34, 124]. Remarkably, [5] first suggested the combination of multiple archives to improve the completeness of web archiving, although at the expense of temporal coherence. Other work also proposed the use of aggregate information from multiple web archives: AlSum et al. showed that the top three web archives in their experiments could provide coverage comparable to their full set of twelve archives [7]. None of these studies, however, studied the use of web archives for security measurements.

Orthogonally, Soska and Christin used archival data to train classifiers for detecting vulnerable websites before they turn malicious [223], while other research teams identified several flaws in how the IA stored archived data, which allowed attackers to compromise the users' view of an archived web page [137, 121]. Though security-related, these are not related to web security measurements.

2.3 Reproducibility

After introducing the foundational concepts of web measurements, we now turn to the topic of reproducibility in scientific research. Researchers should take care to make their studies reproducible, as reproducibility constitutes a fundamental pillar of scientific

trust. It enables independent verification of results, makes self-correction within the scientific community possible, and provides a reliable basis for follow-up research [168].

The term reproducibility is often used broadly and inconsistently in the literature. To address this, several organizations have proposed more precise terminology, including the U.S. National Academies of Sciences, Engineering, and Medicine (NASEM) [168] and the ACM [3]. To clarify the goals and scope of this thesis, we adopt their largely aligned definitions: (*i*) *Repeatability*: the same research team can reliably repeat its own computational results using the same experimental setup; (*ii*) *Reproducibility*: a different research team can obtain the same results using the original authors' artifacts; (*iii*) *Replicability*: a different research team can obtain the same results using newly developed artifacts that implement the same experimental methodology.

While replicability represents the ideal case for scientific findings, in this thesis, we are primarily motivated by transparency aspects and the question of whether independent teams can validate prior methods and have the chance to potentially uncover errors with prior work. Therefore, unless stated otherwise, we focus on *reproducibility*.

2.3.1 A Brief History of Reproducibility in Science

The reproducibility of scientific research has long been a problem. As mentioned in the motivation, the reproducibility crisis, also known as the replicability crisis, has become a defining event in recent scientific history. Concerns about the validity of existing research findings increased in the early 2000s, most prominently with Ioannidis's 2005 essay "Why Most Published Research Findings Are False" [105]. The essay attributes unreproducible results to factors such as small study sizes, biases in study designs, and the publication incentive that discourages the publication of negative findings.

The 2005 essay has sparked a debate that has led to numerous replication efforts across disciplines. In psychology, for example, researchers tried to replicate 13 classical and contemporary effects across 36 samples, failing to replicate 3 (23%) of the effects [125]. Even worse, in cancer research, a study managed to replicate only 6 of 53 "landmark" studies in their field [20]. In economics, replication efforts found that only one-third of the analyzed top-tier studies replicated successfully [42]. These studies represent only the tip of the iceberg in documenting the challenges of replicability and reproducibility across disciplines.

Although often associated with the social, behavioral, and health sciences, reproducibility challenges also affect computer science, including security and privacy research. In the field of usable security, for instance, Klemmer et al. [126] reports that one-third of analyzed papers omit essential information or artifacts, such as questionnaires or software details, needed to replicate findings. Also in more technical domains, such as fuzzing research [205] or machine learning in top-tier security conferences [172], attempts to reproduce prior findings often failed due to poor artifacts. Recent work further highlights pitfalls in the reproducibility of LLM research in the security and privacy domain [71]. Finally, research also shows problems with the reproducibility of web measurement studies, first surveying 117 security and privacy web measurement papers and second demonstrating in an impact study how minimal experimental setup changes can substantially influence results [57].

2.3.2 Reproducibility in Web Measurement

The challenge with reproducibility in web measurements is that the Web itself was initially not designed as a scientific instrument for conducting experiments, but rather as a flexible way to share scientific knowledge and documents. The Web is *ephemeral*: what we observe today is not what we will observe tomorrow, as live websites are routinely subject to changes. The Web is also *erratic*: different vantage points might yield different observations of web pages at the same time, due to modern technologies such as content delivery networks [112]. This means that even when authors tried to do everything correctly, describe their experiment setup in detail, and share all their code, other research teams might not be able to reproduce and validate the findings.

This effect is demonstrated by a number of publications. Ahmad et al. first observed that the choice of a specific web crawler might have a significant impact on web measurements and the inferences drawn from them [4]. Jueckstock et al. similarly showed that specific browser configurations and network access methods might affect web security and privacy measurements [113]. These works pointed out that web measurements are generally hard to reproduce, which motivated additional work by Demir et al. [57]. Their research further confirmed the problem and identified general guidelines to improve the reproducibility of web measurements. However, not only the web crawler but also the location it crawls from plays a role in the measured outcome. In the earlier-mentioned paper [57], the research team presents experiments conducted from various locations. They show that the USA, EU, and Japan note a variance of up to 65% affecting privacy measurements due to different legislation. Similar work shows similar effects for client-side security mechanisms assessing different browsers, locations, and configurations [198] or differing tracking and ad behavior and thus differing JavaScript code [113].

All these papers made the important contribution of identifying relevant shortcomings in published research, however, they only provided limited insights in terms of actionable steps to improve on this limitation. In particular, while guidelines are great for supporting the reproducibility of the measurement methodology, they do not suffice to counter the inherent irreproducibility coming from the ephemeral and erratic nature of the Web. One study [56] even argues that only open datasets can enable reproducibility in web measurements. Motivated by these observations, this dissertation focuses on web archives, which offer a promising countermeasure to these challenges.

2.4 Responsibility

In addition to ensuring reproducibility, researchers must act with moral responsibility to keep public trust. Researchers aim to understand and explain phenomena and make predictions that can inform the decisions of policymakers, society, or individuals. In many cases, such knowledge is produced through empirical studies and experiments. However, as history has shown (e.g., during World War II), not all ends justify the means, and research activities can have serious implications for individuals. For this reason, responsibility must be considered an integral part of conducting research.

Responsibility can be understood as the moral obligation to take ownership of the

foreseeable consequences of one's actions, especially when those actions affect other people's lives, society, or the conditions for future human life [69]. Ethics and ethical frameworks provide guidance for evaluating what counts as responsible by offering principles for judging what is right and what is wrong. Kohno, Acar, and Loh [128] provide a detailed deep dive into the ethical frameworks in security research to which we refer the interested reader. In our work, the aim is not to define what is ultimately right or wrong (a philosophical challenge discussed for hundreds of years), but instead, our goal is to understand what people within and outside the security and privacy community think about the ethics of security and privacy research and web measurement, and about ways to foster ethical thinking within the community. Relevant ethical concepts are therefore introduced and clarified as they arise throughout this dissertation.

To provide more background on this topic, the following sections offer a brief overview of ethics and ethical considerations in security and privacy research, and outline the closely related German legal context relevant to web measurement research.

2.4.1 Ethics in Security and Privacy Research

The importance of research ethics became widely recognized following the Nuremberg trials after World War II. It was a reaction to unethical (and criminal) medical and other human-subject experiments conducted in Nazi Germany [211]. The 1945 Nuremberg Code led to ethical principles and guidelines, such as the Belmont Report [245]. In an attempt to support researchers struggling with these complexities, many universities have established Institutional Review Boards (IRBs), often also known outside the US as Ethical Review Boards (ERBs). IRBs serve as independent committees that evaluate research methods involving human subjects. They review research proposals, provide constructive feedback, and suggest methodological improvements to ensure ethical standards. The significance of IRBs is growing, as many academic conferences now ask for IRB approval for research involving human subjects – a development that has led to new proposals for how such boards should be structured [58].

Still, security research often encounters unique ethical dilemmas that extend beyond direct human interactions [128, 182]. Due to the nature of the digital world, these studies may have broader, indirect implications for individuals, often not recognized by IRBs. For example, the Encore paper [36], which was IRB approved, raised ethical concerns in the review process, as the research method potentially led to browsers of individuals living in suppressive regimes sending requests to censored websites [39, 110]. Similarly, the Hypocrite Commits paper [262], in which researchers introduced vulnerabilities into the Linux kernel, triggered extensive ethical debates and was finally withdrawn from the conference's proceedings [95].

In response to such incidents, leading computer security conferences like the USENIX Security Symposium (USENIX Security) [37] and the IEEE Symposium on Security and Privacy (IEEE S&P) [100] have implemented Research Ethics Committee (REC). Equipped with a more profound understanding of the ethical nuances in security research than IRBs, these committees discuss moral aspects of submitted papers and help clarify ethical questions before making a decision on a paper's acceptance. In contrast to IRBs, RECs only evaluate submitted research and thus do not prevent potentially

harmful research in the first place.

Security and privacy researchers have also tried, even before these prominent cases, to better understand the concerns within the community. The first papers discussing research-ethics concerns and proposing corresponding standards came from the network community [60, 61]. This early work and some informal debates surrounding it led to the Menlo Report [14] and its companion [63]. It is a major framework document promulgating ethical principles for all Information and Communication Technology (ICT) studies, including S&P research, and is still referenced today.

Since the early papers in the network research community, the interest in ethics has further increased within the S&P community. The body of domain-specific work is growing in areas such as network measurements [181, 182], with our work in 3S research [P2], and censorship research [110]. The community has also discussed ethical implications of using datasets of illicit origin, such as password leaks [243]. Ethical considerations emerged even in areas where they may not be self-evident, for example, with regard to cryptographic research [194]. These papers often argue for controlled pre-tests before experimenting on live systems, informing participants in advance and obtaining their consent, practicing mindful and minimized data collection, considering broader societal implications, and engaging in responsible, coordinated vulnerability disclosure.

Another body of work reasons about research ethics through the lens of concrete cases. Kohno, Acar, and Loh [128] use the philosophical construct known as *trolley problem* to discuss ethical dilemmas that can arise in S&P research, pointing out that there is often no consensus on morally acceptable decisions in S&P research. Macnish and van der Ham [143] highlight problems affecting institutional aspects of research ethics, in particular the lack of technical expertise in IRBs. They ground their criticism in two concrete cases: the academic Encore study [36] and the industry-related MedSec case [224]. Both groups advocate for an active discourse regarding ethics in cybersecurity research to raise awareness of these problems.

Ramulu et al. [188] study this discourse by inquiring how ethical reasoning is practiced within the S&P research community. They analyzed how ethical considerations are reported in papers published in 2024 and interviewed authors about research ethics and its challenges. Their findings indicate that ethics in S&P research is often difficult to define, very context-dependent, and commonly guided by the researchers' instincts, prior research, and personal experience. Interviewees further mentioned a lack of formal guidance about what is considered ethical and reported difficulties in identifying relevant stakeholders. Chapter 7 builds on this work. It re-examines and contextualizes several of the findings by empirically identifying a set of perceived ethical goals for the S&P research community. Based on this goal-centered perspective, we further discuss in our interviews new interventions aimed at structuring and advancing ethical discourse and practice within the community.

Zhang et al. [268] analyzes how ethical considerations are reflected in security papers, specifically focusing on discussions of IRB approvals and other references to ethics. Its authors further surveyed researchers in China about their understanding of ethical principles, IRBs or similar boards, and recommendations for ethical considerations in security research. They recommend domain-specific best practices, ethics courses at

universities, and obtaining ethical advice from peer reviewers. However, the paper leaves some open questions, such as how to establish ethics standards within our community and how the community could provide more technical assistance to researchers in terms of ethics. We take up these questions in our interviews and address them in Chapter 7.

While prior work has discussed challenges to ethical research practices in security and privacy research and proposed improvements, domain-specific considerations for the web measurement domain remain underexplored. Moreover, taking a broader perspective, there is limited discussion of an explication of goals that these practices are intended to achieve within the S&P research community. Similarly, little attention has been given to evaluating the ethics-related procedures that conference organizers design and implement.

2.4.2 Ethics Practices at Conferences

In this dissertation, particularly in Chapter 7, we discuss general *ethics-fostering practices* at the leading S&P conferences. As such, we define all activities of individual or collective actors within conference contexts in which they consider, discuss, and evolve the ethical aspects of S&P research.

These practices are implemented in the form of *research ethics-fostering interventions* which constitute specific institutionalized and formalized efforts to modify patterns of actions, specifically behavior or decision-making regarding ethics in our research community. They include formal mechanisms and policies introduced by conferences to promote ethical reflection and accountability, e.g., a section in calls for papers, research ethics committees, or standardized review forms that include ethics-related items.

For review forms, we speak of *ethics-fostering instruments* as concrete operational tools such as evaluation scales, qualitative feedback fields, and ethics information for authors and reviewers.

For simplicity, we use the terms *ethics interventions* and *ethics instruments* throughout this dissertation. When we speak of *ethics* without any further qualifications, this refers to *research ethics*.

2.4.3 Legal Aspects of Web Measurements

Part of conducting responsible research is ensuring that it is carried out within the bounds of the law [14]. While web measurement research on client-side issues (e.g., security headers) can largely be conducted in a legally safe manner using the researchers' controlled browser, for server-side issues, it is more challenging. Prior work has shown that uncertainties in legal regulations can create a *chilling-effect*, where security researchers avoid certain types of research, e. g., 3S, due to the concerns about potential legal repercussions [76]. Indeed, legal action has been taken against researchers in the past [12, 145, 85].

The legal permissibility of 3S research is highly complex, as such scans are typically performed at a large scale over the Internet, potentially affecting many jurisdictions. Since this dissertation includes interviews with legal experts from Germany, we use the German legal framework as an illustrative example to outline potential legal challenges

researchers may encounter when conducting 3S. Germany follows a civil law system rooted in Roman legal tradition, and thus primarily relies on legal codes as its main source of law. As a member state of the EU, its legal system is deeply embedded in EU legislation and judicial institutions.

Criminal Offenses. German criminal law is mainly laid down in the German Criminal Code (Strafgesetzbuch, StGB [79]); Bohlander [27] provides an introduction. Offenses potentially committed via 3S activity include data espionage (Sec. 202a StGB), which is the unauthorized circumvention of access protection to data not intended for the perpetrator and *"specially protected against unauthorized access."* Sec. 202b criminalizes "phishing," defined as *"[the unauthorized interception of data] not intended for [the perpetrator] by technical means from non-public data transmission"* [...]. Sec. 202c penalizes preparatory actions to data espionage and phishing. Due to its wide applicability that also covers the creation of software that could be used for hacking with malicious intent, it has been dubbed the "hacker provision" and widely criticized [45, 80]. The potential for harm caused by 3S research opens the door to offenses related to property damage. Sec. 303a (data manipulation) punishes *"[w]hoever unlawfully deletes, suppresses, renders unusable or alters data [Sec. 202a (2)]"* and 303b (computer sabotage) interference *"with data processing operations [...] of substantial importance"* conducted by means including acts under Sec. 303a (1) or 202a (2) or *"destroying, damaging, rendering unusable, removing or altering a data processing system or a data carrier"*.

Damages. Further, researchers could be held liable under civil law for harm caused through 3S activity. Damages can be issued based on German tort law, as codified in Sec. 823 of the German Civil Code (Bürgerliches Gesetzbuch, BGB [78]). This requires an intentional or negligent unlawful violation of another person's right, including property, or alternatively, a breach of a statute intended to protect another person, such as criminal law provisions.

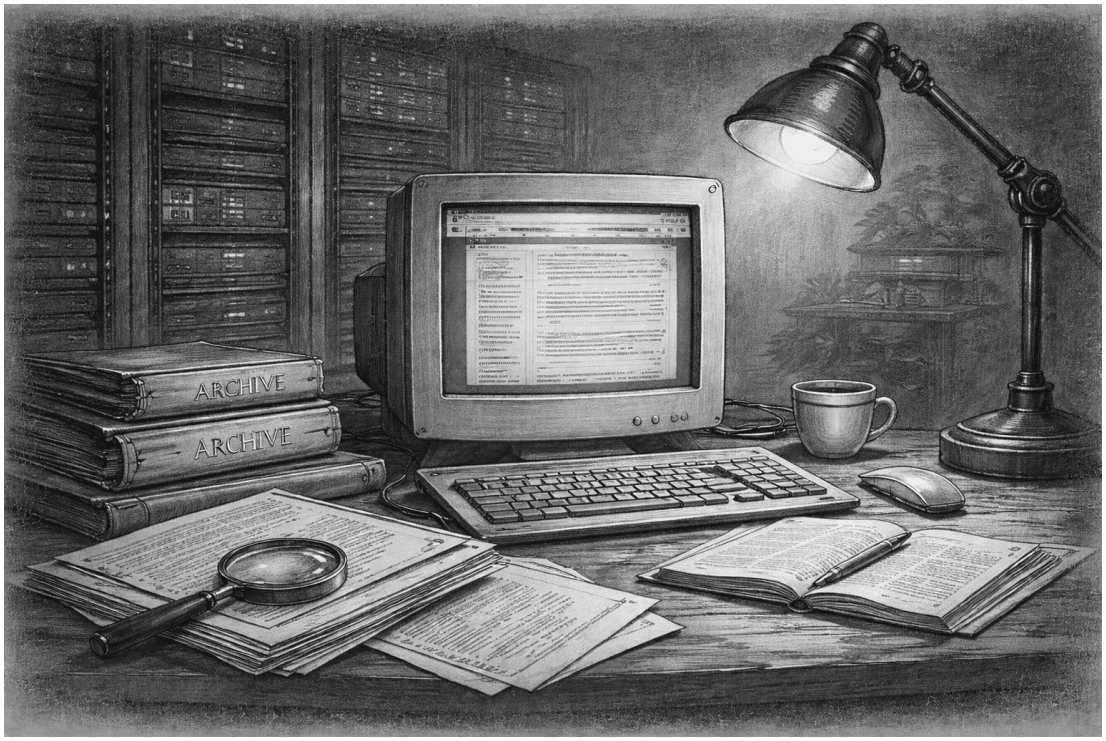
Other. Aside from these fundamental legal risks, server-side research could be at odds with other more specialized areas of law, including data protection law. In Germany, this comprises European laws including the GDPR [70], but also data protection laws at the federal and state level, which, in turn, may constitute additional offenses and legal bases for liability.

Part I

Reproducibility in Web Measurements

3

Evaluating Web Archives for Reproducible Web Security and Privacy Measurements



Contribution of the Author

This chapter is based on "You Call This Archaeology? Evaluating Web Archives for Reproducible Web Security Measurements" [P1]. Stefano Calzavara, Alvise Rabitti, and Ben Stock designed and executed a pre-study for this work. Building on this, Florian Hantke, Stefano Calzavara, Moritz Wilhelm, and Ben Stock designed, (re-)executed further experiments, and analyzed the collected data. Ben Stock mainly implemented experiments evaluating what the best sources of archival data is, Florian Hantke primarily implemented the correctness and live-data analysis, with support from Moritz Wilhelm. Stefano Calzavara and Ben Stock supervised the study. All authors contributed comments and text to individual sections.

3.1 Motivation

As described earlier, web measurements are a popular tool in the security and privacy community for assessing the prevalence of vulnerability classes like XSS, the adoption of state-of-the-art mitigation mechanisms such as CSP, and whether security and privacy best practices like security.txt are gaining traction in the wild. However, experiments are hard to design reproducibly due to the *ephemeral* and *erratic* nature of the Web.

Looking beyond the field of web measurements to other research areas such as machine learning, many fields benefit greatly from shared and reusable datasets. Besides enabling true reproducibility, such datasets encourage incremental improvements (instead of "reinventing the wheel" for each paper) and allow results to be easily, objectively, and accurately compared across papers (instead of results that are specific to particular measurement tools and per-study datasets). But what reusable datasets could we use in web S&P measurement research?

In this chapter, we address this question and explore a solution to design reproducible web measurements – a global archive of web data. Past web security and privacy measurements have made use of web archives to study historical and longitudinal studies to understand the evolution of technologies [170, 138, 232, 197]. We suggest that these archives are already an effective countermeasure to the ephemeral and erratic nature of the Web as they periodically crawl web pages and archive a copy, which is later made available to requesting clients, thus enabling everyone to access the same data and reproduce findings easily.

Although the idea of using web archives for web security measurements has been explored in the past, prior work has only focused on applying them for historical measurements. Yet, we are interested in the correct use of these archives for reproducible measurements. We focus on three leading questions:

- *What are the best sources of archival data available to date, and how can we compare their effectiveness?* Prior security studies are all based on data from the Internet Archive (IA) alone, with the exception of [197], which also used Common Crawl (commoncrawl.org) as a second data source. However, these are not the only available options. The idea of web archives is, in fact, so well established that

the Memento protocol provides a standard access interface to archival data [222, 221] and more than 30 web archives are included in the Memento Time Travel project in 2022.[1] Having multiple data sources may be useful to improve the coverage of archive-based measurements with respect to different domains and historical periods, hence their quality should be rigorously evaluated.

- *Can we trust the correctness of archival data and, in particular, the security conclusions drawn by their analysis?* Prior studies normally assumed archival data to be largely correct, most notably due to the lack of a ground truth. While assuming that archives operate correctly sounds like an acceptable practice, we are not just concerned about archives failing at crawling time, exhibiting buggy behavior, or being actively tampered with, but we are also interested in the bias introduced by the over-reliance on a specific archive. Indeed, each archive operates from a specific vantage point and by means of periodic snapshots. Hence, the points of view of different archives might legitimately be different and biased.

- *Can we leverage recent archival data as an effective substitute for live data for web security measurements?* Although web archives have been used for historical measurements, they have never been used to emulate traditional live measurements. Still, if recent archival data closely matched live data (at least in terms of the enabled security inferences), archive-based measurements could be a viable, reproducible alternative to otherwise ephemeral live measurements. This approach might enable new web measurements that are easily reproducible by design.

We thus make the following contributions in this chapter:

1. We systematically analyze a set of public web archives with respect to the quantity and freshness of their archival data collected from January 2016 to July 2022. Our analysis shows that the IA dominates all competitors, crawling more domains and doing so more frequently. Notably, combining the IA with all the other analyzed archives only gives a negligible additional coverage. Hence, the use of the IA as a single source of archival data is a plausible option (Section 3.2).

2. We assess the correctness of the data available in the IA for historical security measurements. Our analysis is based on two case studies (security headers and JavaScript inclusions), where we identify a few subtleties and pitfalls that may affect the correctness of security inferences (Section 3.3).

3. We investigate the feasibility of performing web security measurements on top of the data from the IA in place of traditional live data, i.e., we assess whether one can make live web measurements reproducible by using archival data, temporally close to the live measurement date. Our experiments show that archival data are crawled often enough to support fine-grained analyses, enable reproducibility, and allow one to draw security inferences close to those of live data, while also revealing important limitations of archive-based approaches (Section 3.4).

[1]The project was taken offline in 2026; however, the data remains available here: `https://web.archive.org/web/20250829170432/https://timetravel.mementoweb.org/`

4. We distill insights and best practices for future archive-based security measurements based on the results of our systematic investigation (Section 3.5).

3.2 Assessment of Existing Web Archives

In this section, we measure the effectiveness of different web archives with respect to the quantity and freshness of their archival data.

3.2.1 Archive Selection

Besides Common Crawl, we consider a public list[2] of web archives supporting the Memento protocol, provided as part of the Memento Time Travel project. We parse the list to collect a set of 64 API endpoints from 38 web archives and we perform a preliminary experiment to identify a set of archives for our study. In particular, we contact all the endpoints to access archival data for the homepage of the top 5,000 domains in the Tranco list [135] downloaded on February 25th, 2022 (ID Z2QWG). We ask for data archived on July 15th, 2022, with a connection timeout of 40 seconds. Note that, since web archives may provide both older and newer content than the requested date in their responses, we are not penalizing archives that did not crawl the Web around July 2022.

At the end of our experiment, we associate to each endpoint a corresponding number of *hits*, i.e., the number of domains for which we successfully received a response from the archive; the more hits we get, the better the archive provides coverage of popular domains. However, this information alone is moot because archival data are not necessarily fresh enough to support useful conclusions. We thus recompute the number of hits for the different archives after setting a *maximum temporal threshold* of six weeks on the collected responses, i.e., if the temporal distance between the requested date and the archival date (available in the `Memento-Datetime` header) exceeds six weeks, we disregard the archival data. We use the term *fresh hits* to refer to hits within the temporal threshold. Note that, since archives may return responses archived on a later date than the requested one, the six weeks threshold actually enforces a rather large window of twelve weeks ($\pm$6 weeks). We argue that any archival data falling out of this window is useless for realistic web measurements as it would be too far away from the requested date.

We report all endpoints which had at least one hit in Table 3.1, along with their number of hits and fresh hits. The table is sorted by the decreasing number of hits and supports several interesting observations. First, although we contacted 38 endpoints, only 13 endpoints had at least one hit, i.e., roughly two-thirds of the tested endpoints are likely outdated, leading to timeouts or other types of failures for all requests sent. As for the remaining endpoints, the four worst-performing ones are operated by libraries and museums, hence they likely index just a small number of resources of interest, i.e., at most 379 domains, amounting to around 8% of the Tranco top 5,000. These archives are certainly not amenable for general web measurements. As for the other nine endpoints, we exclude the End of Term Web Archive because it turned out to be

[2] `http://labs.mementoweb.org/aggregator_config/archivelist.xml` (visited in 2022)

Archive	API Endpoint	Hits	Fresh Hits
End of Term Web Archive	`http://eot.us.archive.org/eot/[DATE]/[URL]`	4,061	3,713
Internet Archive	`https://web.archive.org/web/[DATE]/[URL]`	4,061	3,713
Archive-It	`http://wayback.archive-it.org/all/[DATE]/[URL]`	3,531	2,786
Arquivo.pt	`https://arquivo.pt/wayback/[DATE]mp_/[URL]`	3,412	338
Library of Congress	`http://webarchive.loc.gov/all/[DATE]/[URL]`	2,496	0
Croatian Web Archive	`https://haw.nsk.hr/wayback/[DATE]/[URL]`	2,064	202
Stanford Web Archive	`https://swap.stanford.edu/[DATE]mp_/[URL]`	2,020	96
Archive-It (10702)	`http://wayback.archive-it.org/10702/[DATE]/[URL]`	1,948	0
Icelandic Web Archive	`http://wayback.vefsafn.is/wayback/[DATE]/[URL]`	1,846	986
York University Digital Library	`https://digital.library.yorku.ca/wayback/[DATE]/[URL]`	379	216
Bibliotheksverbund Bayern	`https://langzeitarchivierung.bib-bvb.de/wayback/[DATE]/[URL]`	50	1
Bibliotheque et Archives nationales du Quebec	`https://waext.banq.qc.ca/wayback/[DATE]/[URL]`	42	0
Museum of the Czech Web	`https://wayback.webarchiv.cz/wayback/[DATE]/[URL]`	1	1

Table 3.1: Working endpoints of the different Memento-based web archives, with their corresponding number of hits (for 2022-07-15). The shaded rows represent archives which are further evaluated in this work.

a mirror of the more popular IA and a second endpoint of Archive-It, which indexes just a subset of its resources. The seven archives we consider for further evaluation are shaded in gray in the table. Note that all the shaded archives show an appropriate performance in terms of the number of hits, i.e., they archive data from a significant amount of domains, but some provide just a low number of fresh hits. The reason we do not filter them more aggressively is that even archives with very few fresh hits in 2022 may still be useful for historical measurements performed in previous years.

3.2.2 Coverage and Freshness of Archival Data

Our analysis covers a six-year timeframe from January 15th, 2016, to July 15th, 2022, and is based on periodic archival snapshots of the Tranco top 5,000, taken every twelve weeks. For this experiment, we store all responses in our dataset, but discarded responses lacking the `Memento-Datetime` header because we have no information about their freshness.

For each archive and snapshot, we count the number of fresh hits, i.e., hits for which the archival date is within ± 6 weeks from the requested date. Note that the chosen

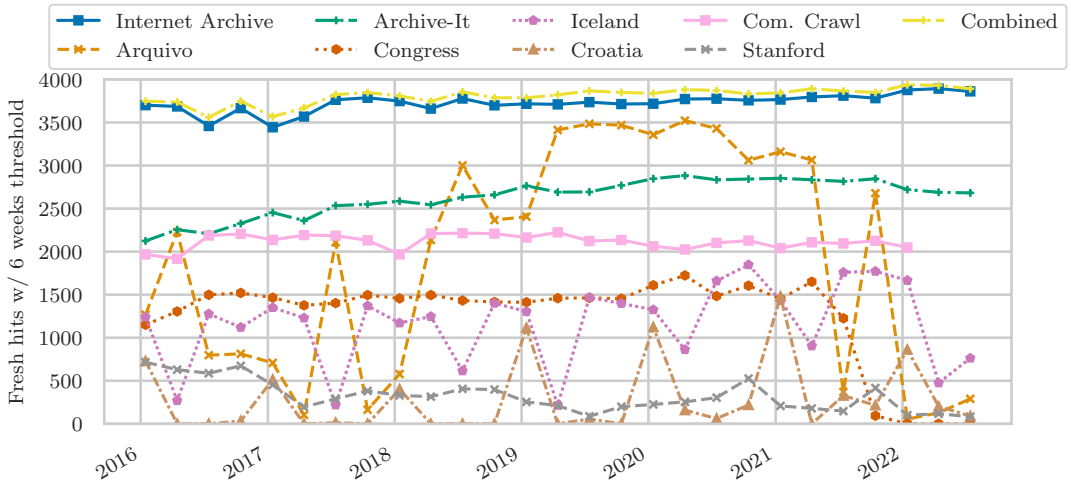


Figure 3.1: Number of fresh hits for different archives

window ensures that all the content fetched from the archives was archived between two consecutive snapshots (twelve weeks away) while avoiding overlaps between snapshots. This is important for meaningful measurements, otherwise a snapshot for June 2020 might, e.g., retrieve data archived in December 2019. Figure 3.1 shows the number of fresh hits for different archives over time. The figure shows that the IA performs much better than all the other archives across all the temporal snapshots that we considered. To further highlight the superiority of the IA, the figure shows in the "Combined" line the number of fresh hits that we would achieve by combining the IA with all the other archives: as we can see, the advantage enabled by the introduction of additional archives is negligible. Finally, the figure also shows that the performance of archives is not always invariant over time and some archives show significantly worse performance in 2022 than in 2016. This is interesting because one might expect that some of the popular domains in the recent Tranco list that we use were not so popular in 2016 and hence were not crawled by the archives. As it turns out, however, some archives just slowed down archiving over time or even stopped doing that entirely, e.g., the Library of Congress does not seem to crawl fresh data anymore. For Common Crawl, the July 2022 snapshot was still not added at the time of our data collection, leading to its line stopping one snapshot earlier.

To better investigate the freshness of archival data, we report in Figure 3.2 a box plot of the temporal distances between the requested dates and the corresponding archival dates observed for fresh hits in our dataset. We consider all snapshots which were available for any of the requested dates. The figure shows that the IA could serve fresh archival data for most of our requests, because its distribution is strongly skewed towards 0, i.e., the requested date and the archival date coincide in most cases, while the other archives suffer from larger fluctuations in general. An exception is Archive-It, which also has a high density around the requested date; this archive covers fewer domains overall (see also Figure 3.1), but for those covered, it provides a high freshness.

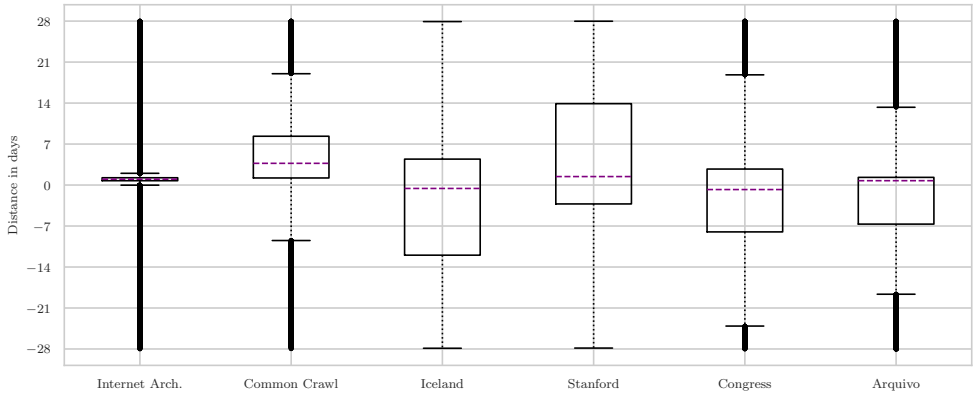


Figure 3.2: Distribution of the observed gaps between requested date and archival date for different archives

3.2.3 Impact of Domain Popularity

Our previous experiment just provides a partial picture of the state of archives because it is based just on the top 5,000 domains of Tranco. To mitigate this bias, we randomly sample 5,000 domains from Tranco (which includes one million domains) and we create a new list of domains with varying levels of popularity. We use stratified random sampling for this task, i.e., we randomly sample 500 domains from each of the ten buckets of 100k domains from the full Tranco list. We then collect two snapshots for the top 5,000 domains of Tranco and the random list of 5,000 domains, one at the beginning (January 15, 2016) and one at the end (July 15, 2022) of our time window. The reason why we consider two snapshots is because we showed that the performance of archives is not invariant over time. We finally compute the number of hits, the number of fresh hits, and the average temporal distance between the requested date and the archival date. As it turns out, the IA is the only archive that is powerful enough to provide reasonable performance on the random set of domains sampled from Tranco because it could provide fresh data for around 2,700 domains in both snapshots (roughly 55%); as for the other archives, the best result was 341 domains (7%). Note that we exclude Common Crawl here given that it does not provide a feasible and cost-efficient way to query for *all* available snapshots, but rather those within a small time window. Hence, it could only ever produce fresh hits.

Most security papers conduct analyses on the top 10k or top 100k [197, 40, 170]. Notably, however, a few studies also focus on the entirety of the top 1M domains [251, 133]. Hence, we investigate the utility of the analyzed archives for the long tail of domains. For that, we analyze the hit rate per sampled 100k bucket in Figure 3.3. The figure shows that the IA is again superior to all its competitors for all buckets. Note that, although Archive-It and Arquivo.pt apparently show good performance on the top 100k domains, their performance drops when focusing on fresh hits. The only archive able to provide fresh hits for a significant number of domains across the Tranco top 1M is again the IA.

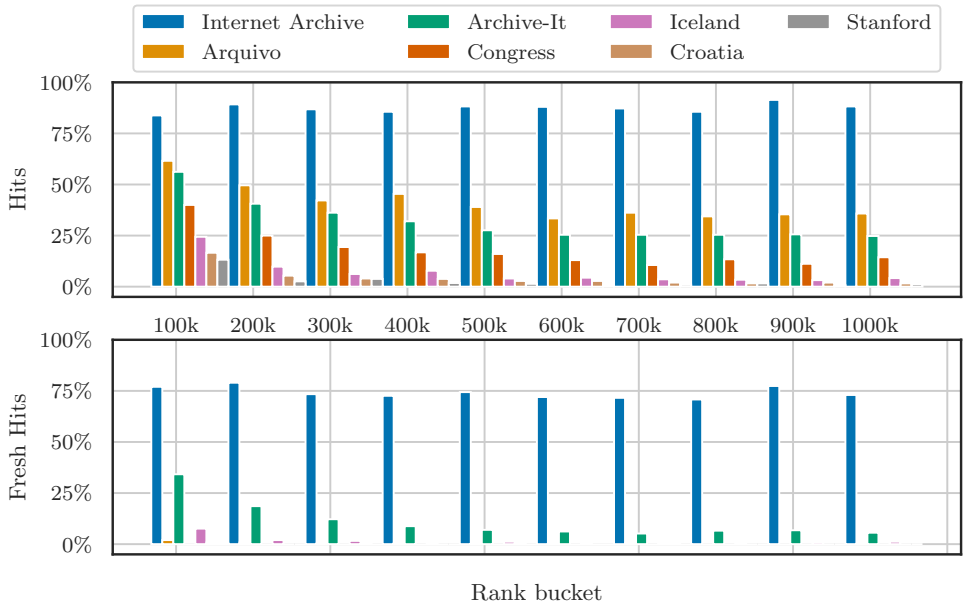


Figure 3.3: Hit rates across popularity buckets

3.3 Historical Security Measurements

We now focus on the correct use of the IA for *historical* security measurements, as performed by prior work which measured the evolution of different web security aspects over time [170, 232, 197]. We just focus on the IA because we showed that it is unquestionably superior to its competitors in terms of both coverage of domains and content freshness.

3.3.1 Methodology

We investigate the quality of the data available in the IA to assess the correctness of historical web security measurements performed on top of them. We are particularly concerned about the bias that might be introduced by the over-reliance on this single data source, since prior research showed that web measurements can be influenced by the choice of a specific vantage point [57, 198]. Our first idea to investigate this was the following: for each snapshot in our dataset, compare the data from the IA against the temporally closest data returned by an independent data source, i.e., any other web archive considered in our prior evaluation. Unfortunately, this simple idea does not work, because the IA is so much better than its competitors that it is impossible to find fresh (±6 weeks from the requested date) independent data for comparison in the vast majority of cases.

We thus leverage the observation that the IA actually aggregates information from multiple data sources (including some of the archives in Table 3.1). For each stored response in the IA, the corresponding source filename is given in the `x-archive-src` header. A source is a collection of multiple responses provided by the same contributor. Feeding this filename to the Metadata API (`https://archive.org/metada`

`ta/<source>`) shows more information about the source, such as the contributor or the crawler name. Since the IA aggregates data from multiple contributors, we can cross-compare its data to reason about the bias coming from the use of specific contributors. In particular, while we cannot directly request and cross-compare data from various archives with the IA as we originally planned, we can reason on how different contributors (including archives) inside the IA might affect our historical view of the Web. Concretely, for each URL u and timestamp t, we collect up to 20 additional nearby snapshots, where with "nearby" we mean within ± 5 days from t; we denote with $N(u, t)$ such set of archived responses, called *neighborhood*. Since all the responses in the same neighborhood are close in time, we expect them to coincide in the vast majority of cases, thus enabling their comparison. We perform both *syntactic* comparisons, where we compare the raw content of responses up to straightforward normalization of dynamic elements (e.g., CSP nonces), and *semantic* comparisons, where we compare the security inferences drawn from the collected data. This way, we can estimate both the quality of the available data and the sensitivity of web measurements with respect to small temporal drifts - an essential property for their generalization.

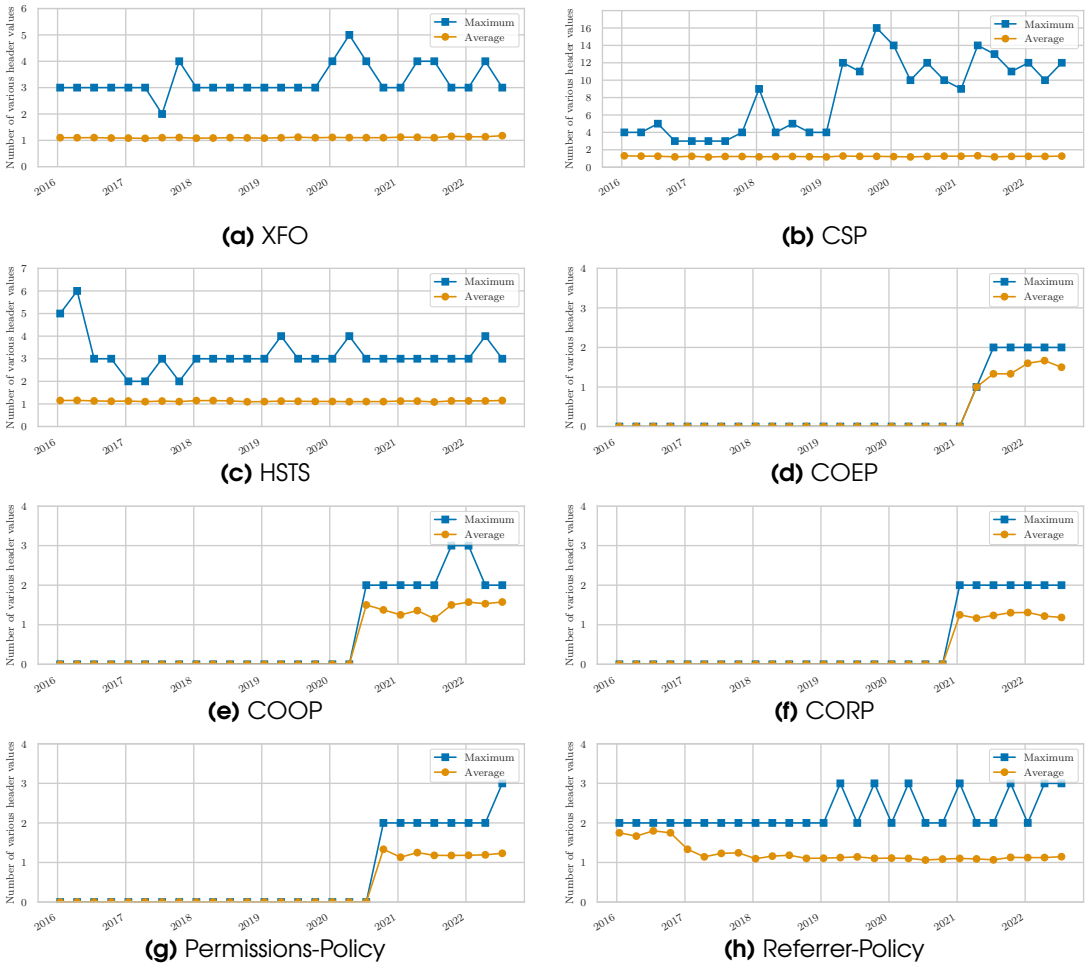
Our investigation is performed on the top 5,000 domains of Tranco, using the archival data of the IA from January 2016 to July 2022, collected as described before (including all response status codes). We focus on two significant case studies: security headers and JavaScript inclusions. These case studies are representative because security headers received considerable attention in the literature [250, 131, 40] and some historical measurements even used archival data [232, 197]. As for JavaScript inclusions, a prior study [170] used the number of remote hosts used for script inclusion as valuable information to reason about JavaScript security because any script included in a web page runs within its origin and inherits the corresponding privileges, hence script inclusion from a single malicious host is enough to completely compromise security. We perform a similar analysis at the *site* level rather than at the host level because the notion of site better captures the concept of a third party [229]. We also investigate the prevalence of trackers among the sites used for script inclusion, similar as performed in prior work [138]. Note that although we borrow from prior work to design case studies worth considering, our focus is different: we are primarily interested in the result of such measurements to reason about the correctness of archival data and historical analyses performed over them. Hence, we do not dive deep into the historical implications of the results themselves, but we rather check how the use of data in the IA may affect (and have affected) such experiments to provide methodological advice.

3.3.2 Security Headers

We first use the IA to investigate the deployment of security headers in the wild from 2016 to 2022.

3.3.2.1 Syntactic Analysis

Our first experiment focuses on the syntactic differences in header values observed within neighborhoods (after header normalization). The graphs in Figure 3.4 plot the average and maximum number of different header configurations observed in neighbor-



(a) XFO **(b)** CSP

(c) HSTS **(d)** COEP

(e) COOP **(f)** CORP

(g) Permissions-Policy **(h)** Referrer-Policy

Figure 3.4: Maximum and average number of different security header configurations within neighborhoods, aggregated per year

hoods over time. We include only neighborhoods that make use of the specific header at least once. Thus, the graphs show that the newer headers have only been appearing regularly for the last three years. They also show that the more recent headers show a less stable trend for the average differences. This can be explained by the low number of neighborhoods that use these headers (Table 3.2), i.e., already a few differences have high impact on the average. Nevertheless, the average number of different header values is close to 1 for all headers, i.e., responses within one neighborhood largely agree w.r.t. the configuration of security headers. This is a positive finding because it means that the bias coming from the use of specific contributors of archival data is limited when measuring this aspect. However, the maximum number of different header values within one neighborhood can be much higher than the average: we found neighborhoods where XFO was configured in 5 different ways, CSP was configured in 16 different ways, and HSTS was configured in 6 different ways. In contrast, the other five headers only show a maximum of three different configurations, likely due to their simpler syntax.

We now investigate which fraction of neighborhoods may observe syntactic differences for the respective security mechanisms. The second group of lines in Table 3.2 shows this: we find that across all headers, at least 10% of the neighborhoods have at least two configurations. The second line highlights that many sites are affected in multiple snapshots since the number of affected sites is much lower than the affected neighborhoods. Notably, though, the vast majority of syntactic differences originate from the respective header not being observed in some snapshots. This has significant implications for works that rely on syntactic changes (e.g., [197]), which should in future work aggregate multiple data points in neighborhoods rather than relying on singular snapshots. The more mature headers, such as XFO, HSTS, and CSP, are more coherently deployed within one neighborhood. Newer mechanisms like COEP and COOP show much higher variance within one neighborhood: in Chapter 3.4, we show that this is related to the fact that these headers are not always collected correctly by the IA, due to specific behavior of Google-owned servers acting as early adopters, thus skewing results.

We discuss a few examples where we observed different header configurations within one neighborhood:

- Some XFO headers are incorrectly populated with different numbers of occurrences of the same directive, e.g., `DENY` vs. `DENY,DENY`. This wrong configuration of the web server may lead to a lack of framing control in legacy browsers [40].
- Some CSPs make use of nonces without putting them within single quotes, as required by the standard, leading to different header values even after our normalization routine (which only operates on correctly quoted nonces). This wrong configuration may void mitigating capabilities for XSS or introduce breakage in web applications.
- Some HSTS headers make use of dynamic `max-age` values, hence different header configurations are archived upon multiple crawls. This dynamic behavior was already observed for live data in the past [198].

Again, a measurement that considers only singular requests could suffer from errors. Instead, we recommend collecting multiple responses for a single snapshot (neighborhoods) and aggregating them to improve robustness, as discussed in this section.

	XFO	CSP	HSTS	COEP
Nbh. setting header at least once	38,417	14,463	31,090	26
- Sites setting header at least once	2,646	1,399	2,464	9
Nbh. w. syntactically different headers	3,870 (10.1%)	2,587 (17.9%)	3,486 (11.2%)	12 (46.2%)
- Affected sites	1,133 (42.1%)	716 (51.2%)	1,044 (42.4%)	6 (66.7%)
- *only because of missing header*	*963 (36.4%)*	*398 (28.4%)*	*832 (33.8%)*	*6 (66.7%)*
Nbh. w. semantically different headers	3,555 (9.3%)	1,210 (8.4%)	522 (1.7%)	9 (34.6%)
- Affected sites	1,077 (40.7%)	377 (26.9%)	210 (8.5%)	5 (55.6%)
- *only because of missing header*	*938 (35.4%)*	*250 (17.9%)*	*141 (5.7%)*	*5 (55.6%)*

(a) XFO, CSP, HSTS, COEP

	COOP	CORP	PP	RP
Nbh. setting header at least once	157	163	1,451	6,780
- Sites setting header at least once	48	47	437	875
Nbh. w. syntactically different headers	72 (45.9%)	36 (22.1%)	272 (18.7%)	704 (10.4%)
- Affected sites	35 (72.9%)	18 (38.3%)	168 (38.4%)	313 (35.8%)
- *only because of missing header*	*29 (60.4%)*	*16 (34.0%)*	*162 (37.1%)*	*275 (31.4%)*
Nbh. w. semantically different headers	3 (1.9%)	16 (9.8%)	-	505 (7.4%)
- Affected sites	2 (4.2%)	4 (8.5%)	-	223 (25.5%)
- *only because of missing header*	*2 (4.2%)*	*3 (6.4%)*	-	*193 (22.1%)*

(b) COOP, CORP, PP, RP

Table 3.2: Neighborhoods (Nbh.) affected by different header configurations

3.3.2.2 Semantic Analysis

We next abstract from purely syntactic differences between header values and focus on measuring the security guarantees offered by the headers. In particular, we consider the following definitions of security:

- XFO: we consider an XFO header to be safe if and only if it provides some restriction on framing, i.e., framing is forbidden on some origins. Concretely, we consider an XFO header safe when it is set to `SAMEORIGIN` or `DENY`. We do not consider `ALLOW-FROM` safe because it is unsupported by modern browsers and was never supported by Chrome in the first place, so its effectiveness in practice is questionable.
- CSP: since CSP has three use cases, we consider different definitions of safety: (*i*) XSS mitigation: the CSP does not suffer from trivial bypasses, e.g., the CSP does not allow script inclusion from any web origin as required by the definition of safe CSP in [41]; (*ii*) Framing control: similar to XFO, we require the policy to provide some restriction on framing, i.e., stop framing on some origins; (*iii*) TLS enforcement: we require the policy to use either the `upgrade-insecure-requests` or `block-all-mixed-content` directives, which have the same effect of forbidding plain HTTP communication.
- HSTS: we consider an HSTS header to be safe if and only if it satisfies the necessary conditions required for inclusion in the HSTS preload list, i.e., the `max-age` directive is set to a duration of at least one year and the `includeSubDomains` option is activated.
- COOP, COEP, and CORP: since COOP is meant to restrict access from cross-

origin openers, we argue that `same-origin` and `same-origin-allow-popups` are deemed safe, while `unsafe-none` is deemed unsafe[3]. Similarly, COEP and CORP are unsafe when set to `unsafe-none` and to `cross-origin`.

- For RP, the primary threat model is to not leak sensitive URL parameters to third parties. Therefore, `unsafe-url` and `no-referrer-when-downgrade` are deemed unsafe, whereas all other values are deemed safe.

Finally, for PP, there is not clear binary choice between safe and unsafe configurations. Therefore, rather than attempt to come up with an arbitrary definition, we omit it from further analysis.

Based on these definitions of safety, responses in $N(u,t)$ can be partitioned in a safe subset $N^+(u,t)$ and an unsafe subset $N^-(u,t)$, and we can estimate how much the responses in $N(u,t)$ agree with respect to safety by computing the *Gini impurity* score as follows:

$$Gini(N(u,t)) = 1 - \left(\frac{|N^+(u,t)|}{|N(u,t)|}\right)^2 - \left(\frac{|N^-(u,t)|}{|N(u,t)|}\right)^2$$

The value of Gini impurity ranges from 0, when all the responses in $N(u,t)$ belong to the same class, to 0.5, when responses are equally split between the two classes. We first estimate the number of cases where the Gini impurity is greater than 0: these cases can compromise the correctness of security measurements because different responses in the same neighborhood enable different security inferences. The results for each security mechanism are shown in the lower three rows of Table 3.2. Across all neighborhoods, we find that between 1.7% (for HSTS) and 34.6% (for COEP[4]) are impure, i.e., they contain observations which have differing security implications. We note, compared to the fraction of impure neighborhoods, that a higher percentage of sites that deployed the respective mechanism at least once are affected. However, the results also show that for the vast majority of sites, this is because they were simply missing the header in at least one snapshot. Considering how many sites are affected (in a potential longitudinal measurement, e.g., [197]), we find that relying on a single snapshot for each site threatens the validity of measurement results. We further study the reasons for these cases in the following section.

3.3.2.3 Attribution of Differences

To investigate why such significant differences occur between neighbors, we identify the response feature f which contributes the most to impurity by computing the *information gain* (reduction of Gini impurity) obtained by splitting $N(u,t)$ in two subsets (i.e., safe and unsafe) based on the value of f. We consider the following features of the responses as potentially important for Gini impurity:

1. *Contributor*: different contributors of archival data might influence the response content due to its vantage point and generic reasons, e.g., bugs in a crawling script;
2. *Status code*: for example, error pages might enforce different security policies than normal web pages;

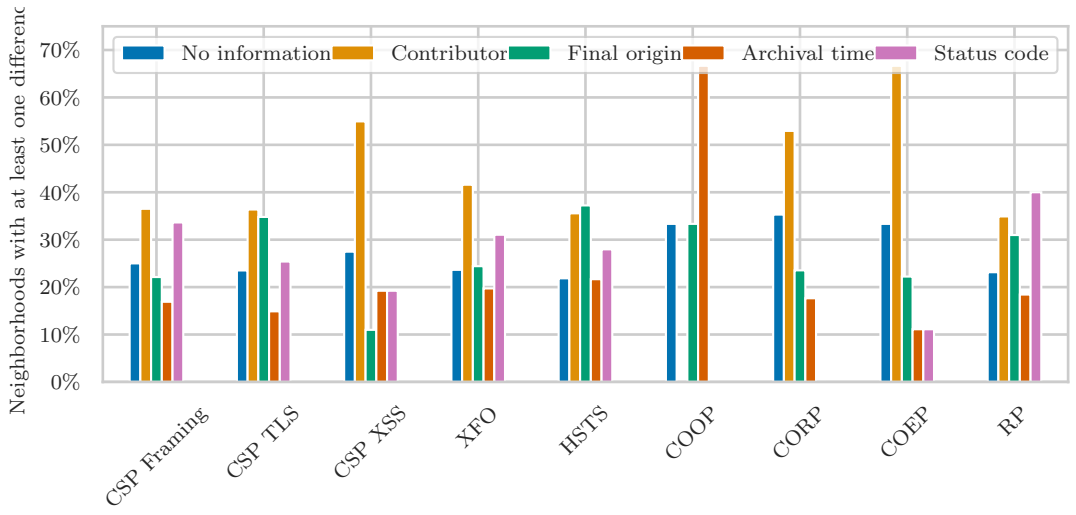
[3] `noopener-allow-popus` was only introduced after this study was conducted.
[4] COEP likely is an outlier due to its low prevalence.

3. *Final origin*: different origins (after redirects) may point to different applications, enforcing different security policies;
4. *Archival time*: changes to security policies might occur even in our short time window of ten days.

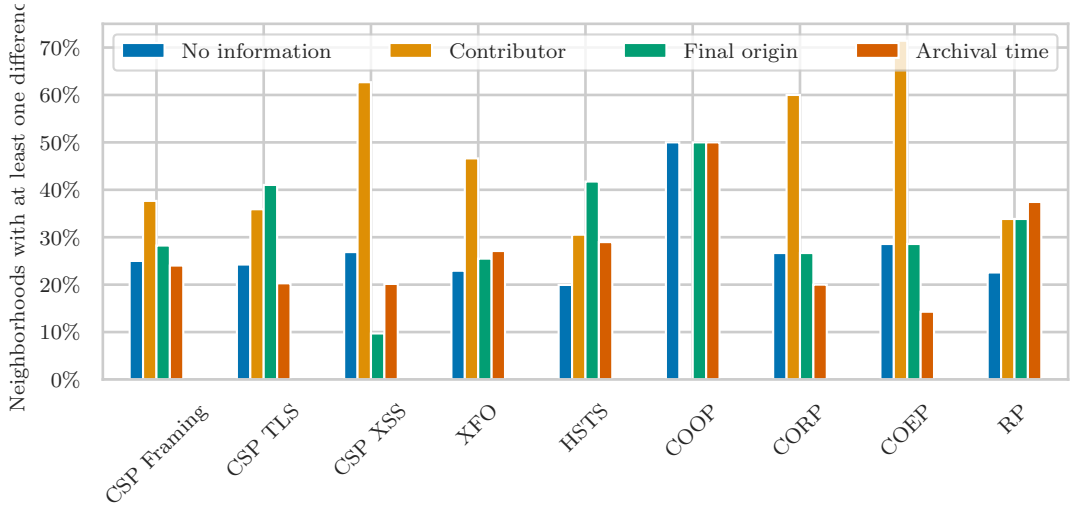
Note that this analysis does not come without caveats. First, multiple features may contribute to explaining the same security difference: for example, different contributors may be redirected to different origins due to their geolocation. We address this issue by considering all these features as equally important when they lead to the same information gain. Moreover, it is also possible that none of the considered features leads to a clear information gain: when none of the features leads to an information gain of at least 0.1, we do not perform any attribution.

Figure 3.5a shows that the *contributor* is the most important feature to explain security differences within neighborhoods for almost all the headers, followed by the *status code*. The *final origin* and the *archival time* play a less significant role in the attribution of security differences (except for HSTS and CSP-TLS). We observe that:

1. The impact of the contributor is significant, yet it can be mitigated by restricting the set of trusted contributors within the web measurement. In our dataset, we observe that many responses (24%) do not bear any information about their contributor, which is set to NULL. Remarkably, the NULL contributor is the one that contributes to explaining most of the contributor-related security differences, i.e., it often disagrees with the other contributors. Web measurements can be made more robust by filtering out the NULL contributor: this does not significantly affect the feasibility of historical analyses, as less than 3% of the considered neighborhoods include only data provided by this contributor.
2. The impact of the status code is significant, yet it can be easily mitigated by focusing just on responses with a specific status code, e.g., in the 2XX class. Figure 3.5b shows a variant of Figure 3.5a where we only keep responses with 2XX status code. As we can see, the picture is not very different, and the contributor is still the most important feature to attribute security differences even when filtering out error pages.
3. The impact of the final origin is often limited, yet it should be zeroed out in correct web measurements. This particularly holds true for measurements related to TLS enforcement (CSP-TLS and HSTS), for which the results show it as having high explanation power. This is to be expected, given that, e.g., HSTS headers have no meaning when sent through HTTP. While different recipes may be used to select the final origin to analyze, e.g., its inclusion in Tranco, it is important to ensure the chosen final origin is always the same across the web measurement because the IA aggregates responses for different final origins under the same URL.
4. The impact of the archival time is often limited in our time window of ten days. We can nevertheless further limit the impact of archival time by making neighborhoods smaller, e.g., using a time window of just five days.



(a) All status codes



(b) Just 2XX status codes

Figure 3.5: Feature importance for the attribution of security differences within neighborhoods (that show at least one difference)

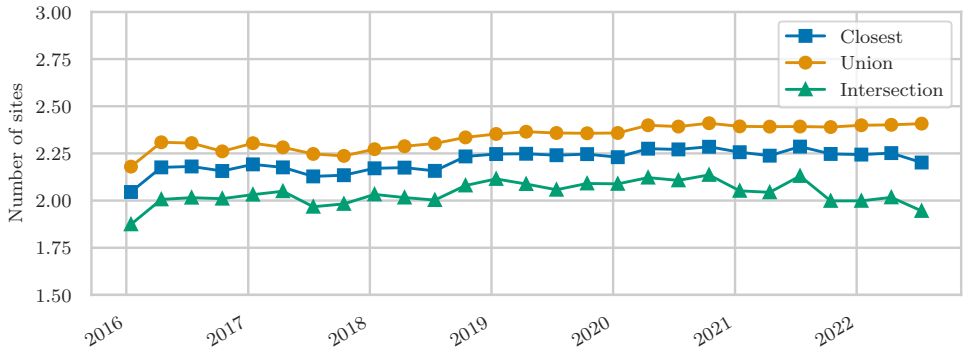


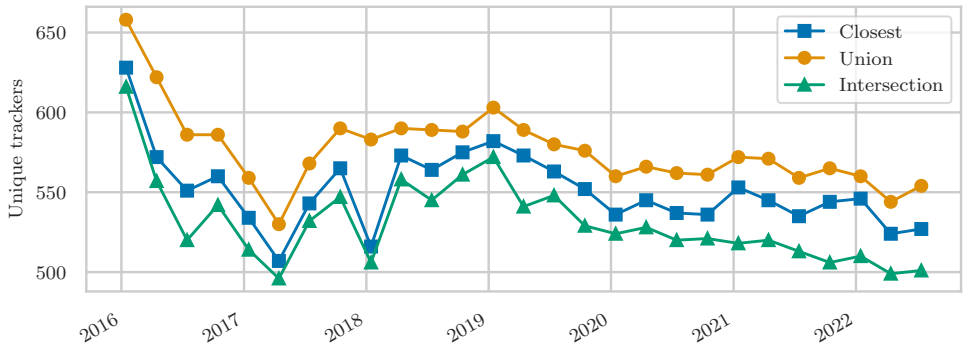
Figure 3.6: Script inclusion statistics for third-party sites

3.3.3 JavaScript Inclusions

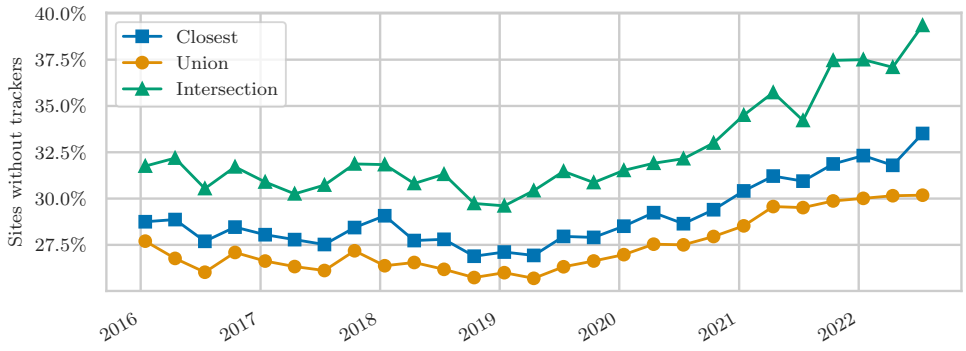
In this section, we use the IA to investigate the state of JavaScript inclusions. We focus on inclusions from remote sites, given their security implications according to SOP. To this end, we parse the static HTML content available in our dataset and collect every URL that is loaded in any script element. Additionally, we use this information to estimate the prevalence of web tracking in the wild (as done in [138]). In particular, we tag a site as a tracker if it belongs to the Disconnect [59] or EasyPrivacy [66] filter lists to estimate how often web trackers are included in popular websites.

In our first experiment, we measure how the average number of remote sites used for script inclusion changed over time. This is shown in Figure 3.6. The plot has three lines: the first one uses the responses which are temporally closest to the analyzed date, while the other two lines show the union of sites detected in each neighborhood (i.e., a site appeared in at least one snapshot within the neighborhood), and the intersection of detected sites (i.e., a site appeared in all snapshots within the neighborhood). This way, we can estimate upper and lower bounds for the sets of sites used for script inclusion, thus estimating the bias coming from the use of the specific data point corresponding to the temporally closest date. All three lines in the plot show that the number of remote sites used for script inclusion did not change a lot over the years. However, although the lines agree on the general trend, the actual number of detected sites may differ. At most, we detect 8% fewer sites when taking the temporally closest response instead of the union of sites, while we detect 12% fewer sites when taking the intersection of sites instead of the temporally closest response. We thus recommend using multiple data points within a neighborhood to make web measurements more robust, e.g., by reporting lower and upper bounds, as discussed in our experiment.

We now zoom in on the prevalence of trackers over time. In particular, we plot how many of the trackers of Disconnect and EasyPrivacy were found at least once in our dataset at different points in time. This is shown in Figure 3.7a. The plot uses the same three types of observations (temporally closest, union, and intersection) as before. The figure shows that the number of unique trackers observed over the years appears to be decreasing. Figure 3.7b additionally shows that the number of websites that do not include any tracker has been slowly increasing over the years. In particular, it shows



(a) Number of unique trackers per date



(b) Percentage of websites without any tracker per date

Figure 3.7: Prevalence of web tracking in the wild based on archival data

	2016		2022	
	static	dyn.	static	dyn.
Average number of trackers	1.95	6.80	1.71	7.23
Average number of remote inclusions	3.41	6.63	3.51	7.04
Unique trackers	465	851	405	899
Websites with trackers (total)	1,497	1,931	1,481	1,855
Websites with trackers (perc.)	73.89%	95.31%	73.10%	91.56%

Table 3.3: Static vs. dynamic crawls of the IA for the set of 2,026 sites available in both 2016 and 2022

the percentage of websites that do not include any tracker *directly* peaked at around 40% in 2022. This trend seems counterintuitive at first glance. However, prior work showed that as of 2021, more than 60% of websites dynamically load third parties [229]. With this in mind, we conjecture that our original focus on static HTML content was introducing a bias in our findings.

We thus implemented a dynamic approach and crawled the archived pages from our previous experiments using Chromium via Playwright. Due to the nature of modern websites (e.g., loading additional scripts during execution), the dynamic analysis significantly increased the number of requests sent to the IA. Instead of one request per website, we now measured more than 100 requests per site on average. This increase in requests for dynamic analyses not only means more use in bandwidth but, more importantly, it also means a drastic increase in the time it takes to request all websites and a significant risk of being rate-limited by the IA. Therefore, to put less load on the IA, we added a 5s sleep after each page request and conducted our experiment just on the oldest and the newest snapshot in our experiments, i.e., January 2016 and July 2022. Despite all the care put into preventing rate-limiting, our crawler still frequently got responses with status code 429 (Too Many Requests). Therefore, we could only compare 2,026 sites. This shows that large-scale analyses of dynamic HTML content are particularly difficult to perform using the IA, which is a significant bottleneck.

Regardless of the bottleneck, we highlight the importance of dynamic approaches for some measurements (e.g., tracking behavior), as the comparison of the static and dynamic crawls in Table 3.3 shows. The first two rows of the table show the average number of remote script inclusions and trackers. In both rows and for both years, the crawler was able to find more trackers with the dynamic approach compared to the static approach. The other rows in the table show a similar picture for the number of unique trackers and the number of websites with trackers. Importantly, the dynamic crawl revealed that the number of websites that include some tracker is much higher than expected according to the static crawl. While both 2016 and 2022 statically show around 73% of sites with trackers, the dynamic ones reveal 95% and 92%. Counterintuitively, the number of sites with a tracker also seems to decline in the dynamic crawl of 2022 (albeit the fluctuation is similar to what was observed in [138]). To confirm the correctness of the IA, we ran another experiment, visiting both live and archived versions of an April 2023 sample of sites. For the overlapping sites which returned a 200 status code in both live and archived versions, 1571 (archived) and 1574 (live) sites contained trackers, thus confirming the accuracy of the IA.

3.4 Live Security Measurements

We now investigate whether archive-based measurements can be a feasible and reproducible alternative to traditional live security and privacy measurements. We just focus on security headers here since our previous experiments showed that JavaScript inclusions need to be studied dynamically, which would put even more load on the IA.

3.4.1 Stability of Live Data

The IA operates by means of periodic snapshots and cannot crawl every website every day. Its view of the Web is potentially coarse-grained: any change taking place between two consecutive snapshots cannot be observed by the IA, i.e., security can be realistically measured through the IA only if it does not change too frequently.

To understand whether the granularity of the snapshots taken by the IA makes it a reasonable vantage point to measure security, we collect live data from the top 20,000 domains of Tranco once per day for fifteen days, and we store them in a local database. We refer to such live data collections as L_i for $1 \leq i \leq 15$ and we use the notation $L_i(d)$ to denote the response from domain d in L_i, if any. We then assess to which extent data collected from live domains are *stable* over our time window, where we say that domain d is stable in the time window if and only if for every two live data collections L_i, L_j we have that $L_i(d)$ and $L_j(d)$ are equivalent from a security perspective. We measure the stability of security headers in two ways. Syntactic stability requires equality over (normalized) header values, while semantic stability just ensures that a safe header configuration does not turn unsafe or vice-versa. Having a high number of stable domains is a necessary condition to ensure that security can be meaningfully measured using periodic snapshots like those taken by the IA.

Figure 3.8 shows how the *syntactic* stability rate of domains changes across the live data collections $L_1, \ldots, L_k$ for increasing values of k. We only compute this rate with respect to the number of domains making use of a given security header at least once within that month. The figure shows that syntactic stability is above 98% after fifteen days for all headers except CSP, while semantic stability (that we do not plot to improve readability) is even higher, ranging above 99%. The header facing the most changes is CSP, the most complex mechanism out of those considered in terms of syntax, however, syntactic stability is above 94% even for CSP. Given this extended period of stability for the vast majority of domains, we conclude that live security headers can realistically be analyzed by means of periodic archival snapshots. Recall from Figure 3.2 that the distance between the requested date and the archival date is normally less than one day for the IA; hence stability over our conservative time window of fifteen days gives strong assurance that security can be measured through archival data. Remarkably, stability does not seem to differ for mature headers such as XFO, CSP, and HSTS and newer headers such as COOP and PP. This is interesting because one might expect that less mature headers undergo more changes, for example, as the result of preliminary testing of header deployment in the wild.

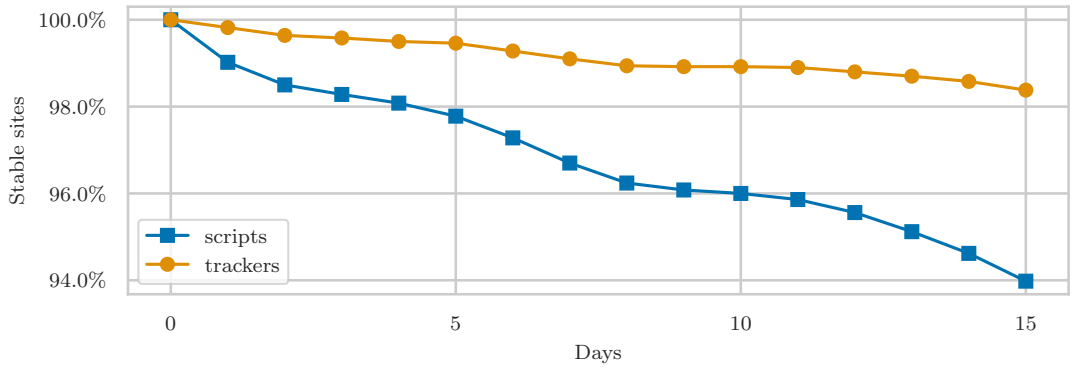


Figure 3.8: Syntactic stability of security headers found in live data between 22/03/2023 and 05/04/2023

3.4.2 Stability of Archival Data

Measurements performed on top of the IA are expected to be reproducible by design; however, reproducibility might break when the IA updates its content. On the one hand, the IA routinely gets updated as the result of its periodic web crawling: even if data is missing for one exact date, a snapshot close after that date might be crawled. On the other hand, requesting data for a particular date (rather than an exact timestamp) might also change the archived content. For example, introducing a snapshot (e.g., from Common Crawl) closer to the requested date may lead to a different conclusion to be drawn with respect to security analyses.

To understand whether the IA can be useful for reproducible measurements, we ask for the same archival data (based on the same requested day r) for the top 20,000 domains of Tranco once per day for fifteen days, and we store them. We only consider fresh data archived within ± 1 days from the requested date r, given our current focus on emulating live measurements. We refer to such live data collections as A_i for $1 \leq i \leq 15$ and we use the notation $A_i(d)$ to denote the (fresh) response from domain d in A_i, if any. We then assess to which extent archival data are *stable* over time, where we say that domain d is stable in the time window if and only if for every two data collections A_i, A_j we have that $A_i(d) = A_j(d)$. This means that three cases can violate stability:

1. Insertions: a snapshot for a previously non-archived domain d has been archived, i.e., there exists i such that $A_i(d)$ is undefined, but $A_{i+1}(d)$ is fresh;
2. Updates: existing archival data for a domain d are replaced by new archival data, i.e., there exists i such that both $A_i(d)$ and $A_{i+1}(d)$ are fresh, but the latter has a more recent timestamp;
3. Deletions: archival data for a domain d become unavailable, i.e., there exists some i such that $A_i(d)$ is fresh, but $A_{i+1}(d)$ returns a response with a status code other than 200.

The latter two operations are particularly dangerous for reproducibility because they imply that the *exact* data used within a web measurement may become unavailable at a later time. Insertions are easier to deal with during experiment design, given that researchers can delay their analysis until the data has become stable, i.e., no new

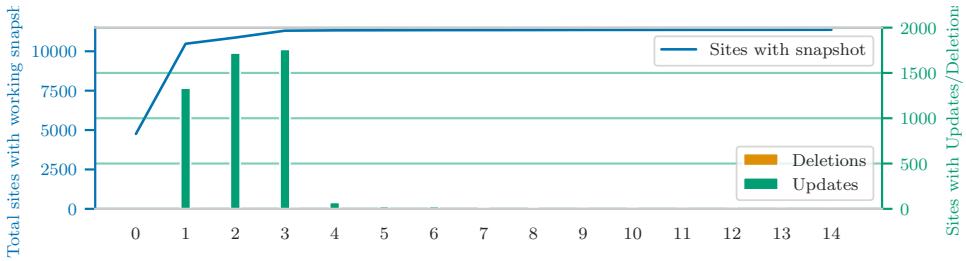


Figure 3.9: Insertions, deletions, and updates of snapshots

snapshots are added closer to the requested date.

Figure 3.9 shows the number of domains with at least one working (i.e., status code 200) snapshot over time (as the blue line, with the left y-axis). Note that given our threshold of ±1 day, we only find data for 11,364 domains. Here, we observe that within the first three days after the requested date, saturation is already reached since there are virtually no snapshots for domains that lacked a previous entry, i.e., no more new insertions. Notably, though, we observe three distinct spikes in updates (shown in the green bars, right-hand y-axis): 1,332 domains, which had been archived before, get a temporarily closer snapshot on day 1. Moreover, after 2 and 3 days, respectively, another 1,720 and 1,760 domains receive updates. This can be attributed to the IA consuming external sources to update its own database. After four days, where a very small number of domains still undergo updates, further updates are negligible in number and become invisible in the figure. Note that the figure also indicates deletions, yet they are always invisible and thus negligible in practice. This leads us to conclude that by consuming external archives, the IA gets fresher data which only in very rare cases leads to a previously archived result becoming unavailable, hence reproducibility is not at stake.

3.4.3 Live Data vs. Archival Data

Having assessed the potential effectiveness of the IA for realistic and reproducible live security analyses, we run a final experiment to establish whether archival data actually matches live data. In particular, we perform both a syntactic and a semantic comparison between archival data and live data, similar to what we performed when measuring the stability of live data. For this, we collect live data of the top 20,000 sites for April 14, 2023 and compare it with archival data for that day, collected on April 17, 2023 (taking into account our previous findings about delayed additions to the IA). We then compare the collected security headers for those URLs with a valid response from both the live crawl and the archived version. Here, we again rely on the tolerance of ±1 day from the requested date. Since prior work [198] showed that web measurements could be affected by the choice of a specific vantage point, we perform three different live data collections from Germany, California, and Australia, respectively, in this experiment.

The results for the security header comparison are shown in Tables 3.4, 3.5, and 3.6. For each vantage point, the table shows the number of sites successfully visited in both the live and archived versions, i.e., both returned status codes 200. At first glance, we

	usage	**syn. diff.**	**sem. diff.**
X-Frame-Options	4,923	198 (4.0%)	190 (3.9%)
Content-Security-Policy	2,640	214 (8.1%)	159 (6.0%)
Strict Transport Security	5,206	378 (7.3%)	367 (7.0%)
Referrer-Policy	1,845	77 (4.2%)	60 (3.3%)
Permissions-Policy	774	228 (29.5%)	228 (29.5%)
Cross-Origin-Opener-Policy	295	207 (70.2%)	205 (69.5%)
Cross-Origin-Resource-Policy	93	33 (35.5%)	25 (26.9%)
Cross-Origin-Embedder-Policy	23	1 (4.3%)	0 (0.0%)
Any header	6,935	602 (8.7%)	536 (7.7%)

Table 3.4: Germany (9,056 sites): Differences between live data and archival data

	usage	**syn. diff.**	**sem. diff.**
X-Frame-Options	4,493	160 (3.6%)	153 (3.4%)
Content-Security-Policy	2,417	162 (6.7%)	118 (4.9%)
Strict Transport Security	4,766	311 (6.5%)	297 (6.2%)
Referrer-Policy	1,741	51 (2.9%)	36 (2.1%)
Permissions-Policy	755	219 (29.0%)	219 (29.0%)
Cross-Origin-Opener-Policy	301	209 (69.4%)	209 (69.4%)
Cross-Origin-Resource-Policy	102	28 (27.5%)	20 (19.6%)
Cross-Origin-Embedder-Policy	22	0 (0.0%)	0 (0.0%)
Any header	6,412	512 (8.0%)	447 (7.0%)

Table 3.5: USA (8,487 sites): Differences between live data and archival data

	usage	**syn. diff.**	**sem. diff.**
X-Frame-Options	4,465	177 (4.0%)	173 (3.9%)
Content-Security-Policy	2,411	188 (7.8%)	140 (5.8%)
Strict Transport Security	4,746	334 (7.0%)	322 (6.8%)
Referrer-Policy	1,742	54 (3.1%)	40 (2.3%)
Permissions-Policy	754	219 (29.0%)	219 (29.0%)
Cross-Origin-Opener-Policy	306	212 (69.3%)	212 (69.3%)
Cross-Origin-Resource-Policy	101	28 (27.7%)	20 (19.8%)
Cross-Origin-Embedder-Policy	22	1 (4.5%)	0 (0.0%)
Any header	6,381	541 (8.5%)	484 (7.6%)

Table 3.6: Australia (8,457 sites): Differences between live data and archival data

observe that relative to the usage of headers, Germany has the highest syntactic and semantic differences, closely followed by Australia. US shows the smallest number of differences. Given the IA is primarily fed from crawls in the US, this is to be expected. For the following discussion, we thus focus on the differences for the US-based crawl.

Notably, when using the IA or crawling live data, the URL initially visited does not necessarily belong to the same registered domain as the final URL. As with our previous analysis, we therefore first looked at cases in which the host of the requested URLs *after* redirects matched between live and archive. Out of the 512 cases of differences in the US analysis, 283 had a differing final origin. As expected, the numbers for Germany and Australia were even higher (344 and 316, respectively) because of geo-specific registered domains. Of the remaining 229 domains, 17 had a much older snapshot in the IA (beyond one day of tolerance). This leaves us with 212 (out of 6,412, i.e., 3.3%) domains for which no obvious explanation exists.

Looking at the table in more detail, it also highlights that more recent additions of security headers, such as Permissions-Policy or COOP, show differences of up to 70% when comparing live with archival data. To understand potential reasons for these (and all other differences), we investigated the sites affected. This revealed that the very large majority of domains deploying these newer security mechanisms are Google domains (`google.tld`), which selectively send the header or are based on user-agent sniffing. Based on this finding, we conduct an additional check against *all* 212 domains with unexplained header differences; making one request against the live site with the User-Agent set to `archive.org_bot` (`https://archive.org/details/archive.org_bot`) and one pretending to be a recent Chrome browser. This revealed that a total of 93 sites employed User-Agent sniffing, simply not returning security headers at all for the IA bot. Of the remaining 119 sites, 34 had a different title (mostly indicating error pages in one case and the actual site in the other), and further four had a similar title, yet the content varied significantly (most likely returned from another server with another configuration). Overall, this leaves us with 81 cases of differences (across *all* headers) that cannot be explained. Relative to the number of sites that deployed any header, this leaves us with 1.26% (81/6,412) without explanation. These findings are in line with recent work from Roth et al. [198], who showed that 127 out of 8,174 sites (1.55%) with any security header in their study had seemingly random differences. Hence, while the IA may show differences of up to 9% compared to live analysis, researchers can a priori design their experiment to limit the impact or filter out bogus data (e.g., caused by User-Agent sniffing). This way, the IA can be used to get similar quality data to live measurements.

3.4.4 Performance Evaluation

The IA can represent a bottleneck for large-scale measurements. In live measurements, requests are spread across different web servers, while in archive-based measurements, all the requests are sent to the IA. This also means that the rate-limiting policy of the IA may hinder the amount of parallelism that would be required to carry out large-scale measurements in a reasonable amount of time. Of course, testing the rate-limiting policy of the IA would be unethical because it would necessarily require sending a high

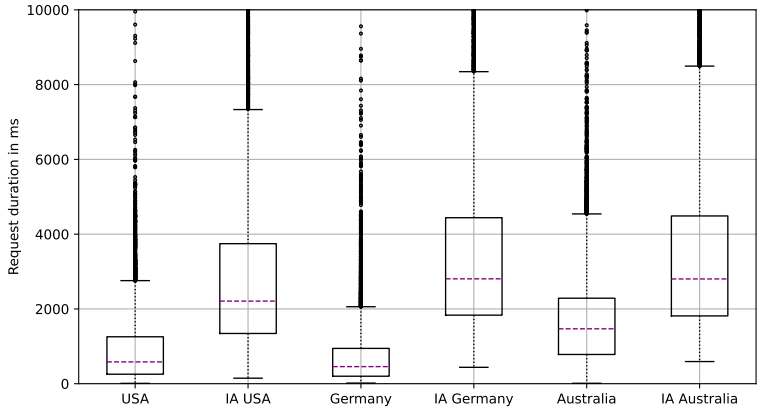


Figure 3.10: Duration for live and archival requests for the three vantage points Germany, USA, and Australia

number of concurrent requests to the IA, which may affect its functionality. We thus limit our investigation to *request duration*, i.e., the amount of time passed between the creation of an HTTP request and the reception of the corresponding HTTP response.

Figure 3.10 shows the distribution of the request time for live measurements and archive-based measurements from the three vantage points for a total of 13,107 domains that could successfully be retrieved (both live and archive) from all vantage points. Requests to the IA are fastest from the US, requiring, on average 3.3s, compared to 3.9s from Australia and 4.0s from Germany. Considering the live requests, which took, on average 0.92, 1.74, and 0.75 seconds, respectively, we note a significant slowdown for single requests. Hence, the slowdown varies from 2.2x (Australia) to 5.3x (Germany). The seemingly lower slowdown for Australia, however, actually relates to the high latency that occurred for live requests, likely because many popular services primarily serve customers from Europe and US (and hence optimize their content distribution accordingly).

3.5 Best Practices for Researchers

Our principled investigation of web archives provides many relevant insights that future research may leverage.

3.5.1 Sources of Archival Data

We experimentally showed that the IA is unquestionably superior to all its competitors for web measurements. Compared to other archives, the IA: (*i*) provides higher coverage of domains in the Tranco top 5,000; (*ii*) provides fresher data on average on the Tranco top 5,000; and (*iii*) is the only archive that can realistically generalize beyond Tranco top 5,000, thus enabling large-scale studies. Moreover, we showed that combining the information in the IA with additional data sources only provides a negligible advantage in practice (Figure 3.1). Our analysis thus shows that using the IA as a single source of archival data is a reasonable choice for archive-based measurements.

3.5.2 Correct Use of Archival Data

We showed that we can mostly trust the data available in the IA. However, leveraging it for research purposes is not as straightforward as one might think. In particular, for syntactic analysis, even for mature headers, measured results might be off for up to 20%. Therefore, our analysis identified several potential pitfalls which may affect the precision of archive-based measurements, and we thus propose a few recipes to improve future studies:

1. The IA aggregates information from multiple sources, hence its point of view of the Web might not reflect any actual vantage point. To improve the robustness of web measurements and mitigate the risk of bias, it is useful to collect neighborhoods of temporally close data rather than single data points, using different aggregation and filtering techniques as discussed in this chapter. Since the majority of incorrect conclusions comes from missing headers (see Table 3.2), checking more snapshots is a simple way to rectify that issue.
2. Different contributors to the IA might disagree with each other. This might come from their specific vantage point or be attributed to other issues, e.g., bugs in their data collection. Restricting the set of trusted contributors might reduce the study's variance and improve its conclusions' robustness. Concretely, we observed that the NULL contributor is the one that is most often in disagreement with other observations.
3. The IA stores responses with different status codes, including error pages. Restricting the focus to 2XX status codes might improve the robustness of web measurements, although we note that some studies may want to also consider error pages. For example, XSS protection should also be applied to error pages to protect the web origin.
4. The same neighborhood of archival data might include responses coming from different origins, yielding responses that are generally incomparable. This is subtler to deal with than status codes because the choice of the final origin to be preferred is not necessarily straightforward. However, it is crucial to always focus on the same origin throughout a longitudinal measurement to avoid "apples to oranges" comparisons between different points in time.
5. The dynamic nature of script inclusions and web tracking complicates archive-based measurements because the static HTML content from the IA only provides limited insights. Unfortunately, the rate-limiting policy of the IA makes dynamic analysis via standard web browsers challenging to perform. Requesting just a subset of resources of interest, e.g., substituting images with placeholders, may help reduce the amount of traffic generated toward the IA.

3.5.3 Archives for Live Measurements

Our investigation showed that the IA can be used to perform security measurements that closely approximate traditional live measurements while being easily reproducible by design. However, archive-based measurements have specific limitations. In particular, our experiments showed that:

1. The IA frequently crawls the most popular domains of Tranco, hence the temporal

distance between the requested date and the archival date even drops to zero for most domains after a few days of waiting. This implies that the archival snapshots cover popular domains roughly on a daily basis and are thus frequent enough to measure security because live data exhibit just small security differences even after a window of fifteen days.

2. When requesting archival content for a given date, the IA does not seem to undergo content changes after a stabilization window of around four days from that date. Hence, archive-based measurements performed after stabilization are largely reproducible. To further mitigate the effect of IA updates, researchers should share the IA URLs from collect data in addition to original request URLs.
3. Archival data closely match live data from multiple vantage points for the vast majority of domains and a number of security headers, including XFO, CSP, and HSTS. This means that archives are almost as good as traditional crawlers for security measurements to measure security trends when following the previously mentioned best practices (Chapter 3.5.2), yet they better support reproducibility.
4. Some domains cannot be analyzed by means of the IA as they implement user agent sniffing, causing different responses for crawlers, which are not representative of live data. While this practice is not particularly widespread, our analysis specifically for COOP showed that by a single large entity (i.e., Google) deploying this mechanism, results could be significantly skewed. This can be accounted for by testing live domains for such behavior. Such domains can then simply be skipped to avoid biases.
5. Archive-based measurements are slower than traditional live measurements. The average overhead of archive-based measurements over live measurements ranges from 2.2x to 5.3x, depending on the vantage point. Also, the rate-limiting policy of the IA may impact the amount of parallelism in the crawling phase. Still, considering mid-to-large scale studies, the increased runtime is offset by the ability to have reproducible measurement results. Notably, though, given the rate-limiting enforced by the IA, it is not feasible to replace full-blown dynamic live crawls given the significant overhead caused by loading vast amounts of resources beyond the static HTML content.

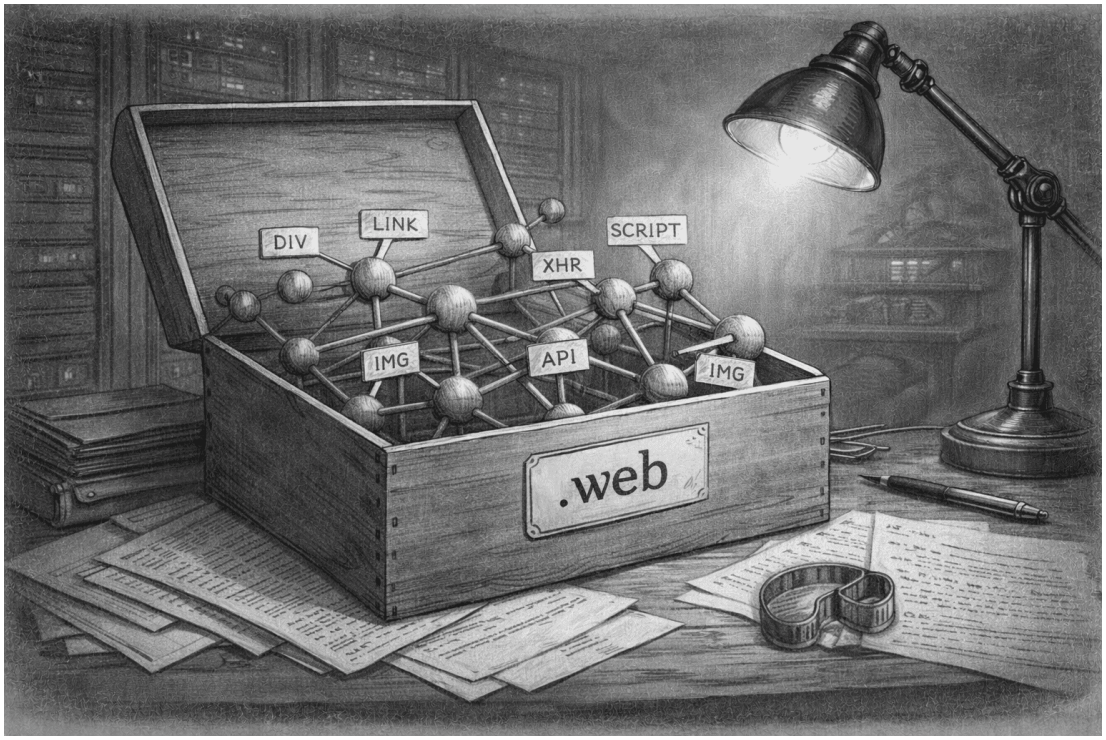
3.6 Summary

In this chapter, we investigated the use of web archives as an approach to conduct reproducible security and privacy measurements. After showing that the Internet Archive is more complete and has fresher data than other alternative archives, we analyzed its correctness by performing historical measurements. We also showed the potential of archive-based measurements as a reproducible alternative to traditional live measurements. Our analysis identified insights and best practices for future archive-based security measurements, providing useful guidance to other web security researchers.

While the Internet Archive is a promising source of archival data, we also showed that dynamic analyses can be difficult to perform on top of it. This brings us to the next chapter (Chapter 4) in which we design and implement a reliable archiving tool and format to overcome dynamic challenges for archival data.

4

Evaluating a Reproducible Web Security and Privacy Measurement Tool



Contribution of the Author

This chapter is based on "Web Execution Bundles: Reproducible, Accurate, and Archivable Web Measurements" [P3]. Florian Hantke, Peter Snyder, Hamed Haddadi, and Ben Stock designed and executed the study. Peter Snyder, together with colleagues at Brave, developed the initial version of PageGraph, the underlying technology we build on in this chapter, and contributed core parts to the collection of tools and techniques in web measurements. Florian Hantke contributed to PageGraph and its surrounding tooling through bug reports and small code contributions. All authors designed the experiments, while Florian Hantke implemented and executed them and led the literature review. Hamed Haddadi and Ben Stock supervised the study. All authors contributed comments and text to individual sections.

4.1 Motivation

As we have seen in the previous chapter (Chapter 3), dynamic content makes it challenging to leverage current web archives as all-purpose tools. Besides this challenge of dynamic content, more fundamental problems exist. For example, consider the following as a demonstrative example to understand how the most common current web measurement tools and web archive formats invite more errors. If a researcher wanted to measure how the Web behaved in 2010, the common way would be to use a browser automation framework like Puppeteer to automate a browser and load a website from a structured HTTP archive such as a WARC file, the structured archiving format used by the Internet Archive. Current implementations of Puppeteer use the current version of Chromium by default and are incompatible with very old browser versions, meaning that websites archived in 2010 will be executed in a 2026 (or later) version browser. However, browser behavior has changed significantly over time, e.g., around 2015 browsers modified how they handle certain "mixed content" resources (see Chapter 2.1.1.4). The effect of this change is that measuring how the Web behaved in 2010 by using modern Chromium versions will cause several subtle but categorical measurement errors. It will under-count JavaScript sub-resource requests while over-counting the number of failed HTTP requests, among other errors. This is just one example.

Current measurement tools can potentially cause many other errors of this kind, as well as other entirely different classes of errors. The lack of accurate, general-purpose web measurement tools and archive formats weakens the field of web security and privacy measurement in at least the following ways: *i.* by encouraging, or sometimes requiring, measurement errors; *ii.* by making it difficult to compare results across measurement studies (since differing results could be introduced by changes in measurement tools, measurement vantage point, etc.); *iii.* by preventing accurate and reproducible historical measurements; and *iv.* by consuming an enormous amount of researcher time on developing similar-but-bespoke measurement tools for each task.

In this chapter, we make a step towards improving the state of security and privacy measurement in two ways. First, by presenting WebREC, a system for accurately

measuring how browsers execute websites, in a manner robust to page circumvention, and covering a broad enough range of browser behaviors (e.g., rendering, network, script execution, etc.) that WebREC can be used without modification for extensive measurements. And second, with *.web*, a novel archival format for recording how a browser fetched, rendered, and executed a website.

More broadly, in this work, we make the following contributions to the goal of accurate and reproducible security and privacy web measurements:

1. A **comprehensive overview of the web measurement and archiving tools** commonly used in research, along with the limitations, why each is insufficient for the goal of generating accurate, reusable, achievable web data (Chapter 4.2);

2. The **design of WebREC and *.web***, a reusable tool and format for measuring and archiving how browsers execute websites accurately and robustly to allow reproducible measurements beyond what current tools and archival formats provide (Chapter 4.3);

3. Multiple **empirical evaluations of WebREC's and *.web*'s accuracy**, compared against existing commonly used tools, finding that WebREC more accurately matches baseline counts of events (Chapter 4.4);

4. An **empirical measurement of WebREC's generality**, by surveying recent security and privacy web measurement papers and evaluating which could have been performed with WebREC directly, or *.web* archives, avoiding the need to create new, custom measurement tools (Chapter 4.5);

4.2 Tools and Techniques for Web Measurements

In this section, we provide a brief overview of existing tools, techniques, and formats commonly used in web security and privacy measurements, based on a literature review of top-tier papers, such as those listed in crawling SoK and survey papers [225, 4]. Our goal is to be demonstrative of the common approaches used in research and discuss why these are insufficient for the replication, accuracy, and iterative needs of scientific research. We then follow by describing common formats used for archiving websites, so that measurements can be done retrospectively, after websites have been changed or are no longer available on the Web. We highlight each approach's strengths and its limitations that make it insufficient as an accurate, general-purpose archiving format.

4.2.1 Existing Web Measurement Tools

We first describe four general categories of tools commonly used in web measurement. Based on our literature research, we find that all web measurement tools used in top security and privacy papers fall into one of these categories.

4.2.1.1 Recording HTTP Requests

The simplest, and earliest, approach used in web measurement research is to record the HTTP communication for each page being measured. In this approach, researchers select a tool for issuing an HTTP request to a URL (e.g., curl [51], wget [82], among many others), use the tool to request the HTML document for the webpage being measured, and record the outgoing HTTP request and the server's HTTP response. Numerous examples can be found in literature [28, 40, 193, 104].

This approach can be extended to use the same tools to record the HTTP communication for sub-resources (e.g., images, or script files) referenced in the HTML of the webpage, either by searching for URL patterns or parsing the HTML.

The defining feature of this approach is that tools are being used to understand or approximate how users experience the Web, but without actually incorporating the main tool people use to interact with the Web; web browsers.

Limitations: Measuring the Web based on raw HTTP request data has important benefits; hundreds of instances of `curl` can be run in the amount of resources needed to load a single webpage in Chrome. However, this approach also comes with serious limitations; Evaluating HTTP and HTML without using a web browser will give, an extremely *lossy* approximation of peoples' web experience (as we have seen in Chapter 3), making it unsuited for answering many web-related research questions.

As example, web browsers use complicated rules for deciding which sub-resources to fetch, affected by many factors like a page's *Content Security Policy* [260], `preload` [255] instructions given on previous pages, among others. Concretely, consider a dynamically added image element; the image is typically fetched when the element's `src` attribute is set, regardless of whether it has been appended to the DOM. These edge cases require an extreme amount of domain-specific expertise, leading to accidental but *preventable* errors.

In short, tools in this category focus on HTTP communications but do not consider how web browsers parse and execute those websites. Therefore, they are only suited to answer questions about server responses, but not accurate enough to measure how the Web is experienced by most people.

4.2.1.2 Browser Instrumentation

A second approach is to leverage the browser itself, either through automation APIs (i.e., WebDriver [258] and DevTools [47]) and popular libraries for interacting with these APIs (e.g., Puppeteer [83]), or through browser extensions. These approaches are used in a variety of web research [186, 203, 55, 190] and account for the limitations of the previously discussed approach by measuring websites as executed by browsers.

All of these browser instrumentation techniques are able to measure how a website is rendered, cookies and other values are stored, and execute and inspect JavaScript in the page by injecting proxies into global JavaScript structures. Some tools also measure network requests made by the page.

Limitations: While these approaches are extremely common in web security and privacy measurement research, they have important limitations, making them unsuited for measuring certain phenomena researchers investigate. Broadly speaking, browser

instrumentation tools lack APIs to directly measure many page behaviors, requiring researchers to try and write code that replicates browser behavior. In the best of cases, this invites *preventable* implementation errors; in many worst cases, measurement errors are unavoidable or *unpreventable*.

For example, browser-instrumentation-based measurement tools are not able to observe JS execution directly, requiring researchers to rely on indirect best-effort techniques (e.g., modifying the execution environment). These techniques lead to incomplete or incorrect results for a range of reasons. Measurements that rely on modifying a script's execution environment fail when page scripts regain access to unmodified JS prototypes (e.g., by storing a reference to the original prototype early during page execution), something extremely difficult to prevent and a pattern utilized by real-world advertising libraries or anti-debugging techniques [166]. Furthermore, measurements relying on interposing of non-configurable global JS elements cannot capture the interaction with these (read-only) Web API properties, like `window.location`, making errors *unpreventable* and leading to a lack of *comprehensiveness* with current tools.

These are just some examples of phenomena that browser-instrumentation-based measurement tools have difficulties in *accurately* measuring. While suitable for functional tests (the task these tools were initially designed for [208]), these tools have significant limitations preventing them from being the *general* purpose solutions in web measurement research.

4.2.1.3 Browser Orchestration

A third category of web measurement tools builds on browser-instrumentation tools, relying on the same browser capabilities and intervention points but creating reusable and improvable instrumentation code for consistent measurement tasks. This allows a common body of code to improve over time by benefiting from the collective knowledge of multiple researchers. In short, this third category of web measurement tools improves measurement accuracy by pushing the level of abstract up another layer, allowing researchers to conduct measurements more easily and accurately without needing to personally understand the peculiarities of the web platform.

The most popular example of this category of web measurement tool is *OpenWPM* [67], used by studies all across the community [28, 107, 41, 267]. In addition to the main benefit of providing an open and reusable code base that researchers can use and improve, *OpenWPM* also has additional benefits. First, *OpenWPM* provides standardized inputs and outputs for conducting measurements, further aiding replication. And second, it also includes scheduling functionality, making it easier to measure websites in parallel across different machines.

Limitations: *OpenWPM* relies on the same underlying systems and capabilities as the previously discussed browser-instrumentation-based tools and so shares many of the same limitations. While *OpenWPM's* reusable, iteratively-improved code base reduces the chances that researchers will make *preventable errors*, *OpenWPM* and similar systems fundamentally cannot address the *unpreventable errors* caused by the limitations of the underlying capabilities OpenWPM relies on (see limitations in Chapter 4.2.1.2), e.g., not *accurate* enough to *comprehensively* measure all Web API properties.

4.2.1.4 Browser Modification

A fourth approach to web measurement is to try and avoid the limitations of the APIs available to *browser instrumentation* approaches by directly adding instrumentation to the browser's code base, for example the DOM construction. Since all popular browsers are (in part or in full) open source [240, 155, 248], researchers can accurately measure any browser functionality by modifying their code.

This browser-modification approach has been used for a wide range of measurement tasks, generally to measure lower-level browser behaviors that are not directly visible to page-executed JS or the APIs available to the Puppeteer libraries. Researchers have published research depending on browser modifications to log script compilation and fine-grained JS execution [111], detect XSS attacks [235], the behavior of injected style sheets [134], and to perform taint analysis in JS code [123], to measure the behavior of client-side sanitizing libraries among many other examples.

Limitations: Measurement tools that modify how browsers are implemented provide researchers with powerful measurement capabilities. However, this flexibility imposes a wide range of limitations on these tools. Primarily, the correct implementation of such tools requires an extremely high amount of domain knowledge in the implementation and esoteric standardized behaviors of modern browsers, making *preventable* errors easy when instrumenting target behaviors. And while some of these errors may be obvious to researchers, e.g., build errors, other errors can be extremely subtle, and require domain expertise to catch, e.g., miss-attribution of edge cases.

As a demonstrative example of the difficulties of correctly implementing browser-modification-based measurement tools, consider the task of modifying Chromium to log access to browser APIs (the goal of several projects, such as [106, 111]). Measuring which script is calling which API *completely and correctly* requires understanding (1) the low-level implementation details, (e.g., the differences between a V8 `Isolate`, a V8 `Context`, a V8 `ScriptState`, and a blink `ExecutionContext`) (2) subtleties in the browser application model (e.g., the difference between the calling and receiving execution contexts), and (3) a comprehensive knowledge of all code interacting with these details. This task is not uniquely difficult, and comparable difficulties exist for instrumenting the parsing behaviors, JS ordering, among others.

We emphasize that these *are not* criticisms of existing research tools, which often aim for a general understanding of one phenomenon rather than complete correctness. Instead, we note these difficulties to show how current tools are not well suited for the tasks of a *generalized* web measurement tools, i.e., they are not reusable for everyone for a wide range of measurement tasks without modifications.

4.2.2 Existing Web Archiving Formats

We next provide a brief overview of common archiving formats used in web security and privacy measurement and highlight their limitations.

4.2.2.1 URLs and Raw Resources

The simplest archive form generated by web crawls is a list of visited pages' URLs, allowing revisits and re-measurements in the future (e.g., the Tranco list [135]). Lists like this are used to analyze phenomena like typosquatting [237] or similar.

Other archival datasets might include the actual resources described by the URL at measurement time, such as an image dataset containing both the URLs and the image's content. Including both makes the dataset more robust, since, even if the hosting server no longer returns the resource from the URL, the resource of interest is still available in the dataset.

Limitations: Such datasets are popular because they are easy to generate and to use. However, this approach has profound limitations, making it poorly suited as a *general* approach to web archiving. Most fundamentally, such datasets lack many other resources and information present when the website was loaded and rendered (e.g., HTTP headers), limiting the ability to ask new questions about old data.

4.2.2.2 HTTP Archives with Metadata

The other common form of web archiving is to record the HTTP headers and bodies of all requests and responses during the page's execution, along with metadata like timestamps. Frequently used formats are the web Archive format (WARC), used by the Internet Archive, and the HTTP Archive format (HAR), e.g., exportable from many browsers' devtools. Both formats record largely the same kinds of information, though with minor differences, such as WARC de-duplicates requests while HAR does not.

Sites archived in these formats record the communication while executing a target web page. As we presented earlier, researchers can use such archives for reproducible measurements or historical analysis [186, 197, 138, 232]. They replay responses by loading the initial request and the loaded sub-requests from the archive instead of from the network. Effectively, researchers attempt to re-run the web site from the archive to approximate how the site operated at measurement time.

Limitations: These archives are very useful for measurements that study the resources loaded by a page, as sub-resources are archived exactly. While extremely common in web S&P research, these approaches, however, are significantly *inaccurate* when trying to replicate how websites behaved during execution.

There are multiple reasons why these archive formats do not accurately enable future researchers to replicate how archived websites behaved. For example, the web platform involves many sources of non-determinism (e.g., the current time, random values, available system resources, etc.) which can cause websites to behave differently on re-execution – an *unpreventable* issue when this behavior was not captured.

Similarly, changes in browsers and archiving tool behavior over time further introduces errors and inaccuracies. Browsers' behavior at *archiving* time may differ significantly from how browsers behave at *replay* time, introducing entire categories of potential inaccuracies. For example, consider the aforementioned change of the *mixed content* policy. Modern browsers block insecure JS files while browsers ten years ago had no such restrictions [259]. This change means replaying such archives in modern web browsers will cause categorical inaccuracies, making it seem like users historically

encountered far more blocked requests than they actually did.

This change is just one example of how such archives can lead to systematic errors in historical web measurement. Differences in the *archiving* and the *replaying* environments in any of the following will cause measurement errors: differences in i. browser (e.g., Firefox v.s. Chrome), ii. browser version, iii. operating system, iv. updates in available Web APIs, v. *headless* or *headed* browser modes, and vi. browser language and locale preferences, among many others. These errors might be *preventable* when the exact infrastructure is replicatable, in many cases it is not [57].

4.2.3 Summary: Limitations and Requirements

Finally, we note the overall challenges facing researchers attempting to conduct accurate, replicable, shareable web research, given the limitations in existing web measurement and archiving tools, leading to requirements for a new tool.

Prohibitive Required Domain Expertise: The vast majority of non-trivial measurements of web behaviors require a large amount of domain expertise. *Accurate* measurements demand understanding and often re-implementing various browser behavior, challenging even for the industry teams implementing browsers, let alone academic web researchers.

In the best of cases, the expertise needed for *accurate* measurement means researchers must familiarize themselves with esoteric browser concepts to *accurately* approximate how users experience the Web. In the worst of cases, the high level of required expertise results in incorrect results.

Redundancy and Correctness: The current web measurement field lacks tools that are both *general* for use across a wide range of measurements, and implemented in a manner that yields *accurate* results, i.e., avoid aforementioned *comprehensiveness* limitations (Chapter 4.2.1.2). Instead, projects typically implement a new *instrumented browser*, typically without correctness tests beyond checks by the researchers themselves. The systematic result of such redundancy is a *worst of all worlds* situation: a lot of time spent (i.e., research teams re-implementing similar tools) creating tools of unknown correctness (i.e., often no correctness verification).

Limited Dataset Sharing: Despite the enormous amount of web measurement work conducted and published, our field lacks *general* datasets for *comprehensive* measurement tasks. The closest our field has are URL lists or collections of archive files, both with discussed limitations. The lack of general datasets of website behavior hinders iterative improvements that we typically see with datasets in other fields like ML.

Tool Requirements: As a result, web measurements result in independent executions of (often) different websites in independently implemented measurement tools, making it difficult to confidently compare results across papers.

To address this, a new tool is needed, especially to handle the *unpreventable* errors from Web API restrictions and replicability problems with dynamic behavior. We developed the following requirements for such a tool. First, it must be *comprehensive* and support a wide range of measurements out of the box, including JS executions, without requiring specialized domain expertise. Second, it must be *accurate*, also when handling JS observations. And third, the tool must be *general*, producing reusable

archives that go beyond only storing a site's resources but also provide all the executions to enable accurate replication of experiments.

4.3 WebREC and *.web* Archives

In this section, we introduce *Web Execution Bundle RECorder* (WebREC), a tool for accurately measuring a wide range of browser behaviors during the execution of a webpage, and *.web*, a three-part archive bundle format that WebREC uses for recording its observations. The following subsections detail the designs of WebREC and *.web* and how they overcome the limitations of existing tools described in Chapter 4.2.

4.3.1 Overview and Goals

WebREC is designed to be *accurate*, *general*, and *comprehensive*, i.e., it can capture a wide range of browser behaviors and events, making it suitable for a large fraction of security and privacy web measurements without the need for modification; we expect that data necessary for diverse studies is a subset of what WebREC currently measures.

WebREC records its measurements as a single *.web* archive, which is a bundle of three child files, described in more detail in the following subsections. The format of these files is consistent and well structured so that common tooling can be used to query the page behaviors and events recorded in each *.web* archive.

We note that *.web* archives contain all observed events that occurred during WebREC's execution of the page and not a subset of behaviors specified at *measurement time*. This has the cost of making *.web* larger than needed for the immediate measurement task but brings two corresponding benefits.

The first benefit of WebREC producing *comprehensive* recordings of page behavior is that *.web* archives are ideal for long-term, cross-purpose archiving. These rich and comprehensive archives allow researchers to ask a wide range of questions about historical data, including questions that might have been unanticipated at *archiving time*. We note that this makes *.web* archives very different from the existing archive formats discussed in Section 4.2.2, where datasets are generally narrowly focused on a particular research question. Moreover, no replaying of requests is needed, a limitation that we showed with WARC and the IA in Chapter 3.

And second, *.web* is *general*, allowing it to generate archives to support reproducible studies beyond what existing common measurement tools and archive formats allow. Consider the demonstrative "mixed content" example from earlier, for instance. Because WebREC directly records how the browser behaved when executing a page, it allows future researchers to accurately measure how the website (including mixed content) was experienced by past users. This is very different from existing archive formats, which record *what the browser fetched* instead of *how the browser behaved*. Only knowing what resources were fetched for past websites hinders historical accuracy and reproducibility by making it difficult-to-impossible to know if differing results are due to *endogenous differences in measurement correctness* (e.g., detecting an error in the original measurement's approach like in Chapter 4.4.5.2) or *exogenous differences in page re-execution* (e.g., browser non-determinism as discussed in Chapter 4.2.3).

In summary, WebREC is designed to be a tool for *general*, *accurate*, and *comprehensive* web security and privacy measurements, and produces *.web* archives designed to enable accurate, reusable, general data archiving to aid reproducibility of web measurement.

4.3.2 WebREC Design

We implement WebREC as a modified Chromium browser (on top of Brave Browser). As explained in Chapter 4.2.1.2, this abstraction level is needed to ensure *accurate*, high-fidelity insights into website executions. Chromium's market share of over 70% [228], makes it a very *general* and long-term solution. Naturally, this decision comes with the limitation that WebREC is browser-dependent. However, we do not focus on cross-browser differences but rather on capturing *accurate* and *comprehensive* web behavior. Alternatively, supporting all major browsers would add significant complexity and maintenance challenges.

WebREC consists of three parts: (1) a system for determining *which* page(s) is measured, (2) a system for measuring *what* resources are fetched during page execution, and (3) a system for measuring *how* the browser behaved while executing the page. Each capability is briefly described below.

First, a researcher directs WebREC to measure a webpage in one of two ways: either first, by using the existing, well-known, and previously discussed Puppeteer library to pragmatically navigate WebREC to a page to measure, or second, by using WebREC in an interactive session, where the researcher uses the browser as a user typically would, but with WebREC recording all relevant page behaviors.

Second, WebREC measures what resources are fetched during page execution using Chromium's built-in *DevTools* protocol, the same protocol used by Chromium's debugging tools and Puppeteer to automate Chromium instances.

Third, WebREC records browser behavior during page execution by instrumenting Blink, the open-source system used for parsing HTML, presenting web pages, and handling the interaction between JS and browser APIs used in Chromium. Our approach builds on an existing Chromium-based measurement system called PageGraph, which has been used in prior work [46, 216, 215]. PageGraph *comprehensively* tracks and attributes all network requests, DOM changes, and denoted WebAPI and JS-builtin calls, building an internal graph of annotated *actors*, *actees*, and *actions*. WebREC extends PageGraph in a number of ways, including handling and recording redirection in HTTP requests and recording the arguments to (and structured responses from) WebAPIs, among other changes needed to support the high-fidelity and retrospective use cases WebREC aims to target.

Finally, we note several attributes of WebREC's construction that better ensure accuracy and help to avoid some of the correctness issues that affect some ad-hoc, or per-project, measurement tools. For example, WebREC is validated against the cross-browser Web Platform Tests [261] to ensure stability and coverage (i.e., that WebREC can handle and record the enormous range of browser behaviors used by modern websites). Additionally, WebREC is kept current with every major Chromium release, ensuring that it accurately records the Web as executed in a then-current

browser (see Chapter 4.6.5).

In short, WebREC is built to minimize (or avoid) the limitations of other browser-modification web measurement tools.

4.3.3 *.web* Archives

Finally, we discuss the format and contents of the *.web* files WebREC produces when measuring a page's execution.

A *.web* archive consists of three files. First, it contains a screenshot of the measured website at the end of the measurement period (i.e., when WebREC stops measuring page behaviors and begins serializing it into a *.web* archive). This screenshot is captured using the Puppeteer library, similar to other web measurement tools.

Second, each *.web* archive includes a copy of resources fetched during the page's execution. This archive is also collected using Puppeteer and is encoded as a HAR archive, again, in a manner similar to existing measurement tools.

Third, and most novel, each *.web* archive includes a description of all page behaviors measured during the page's execution. Rather than using existing archiving formats like HAR or WARC, we adopted PageGraph's graph-based format to represent element dependencies and action flows as a directed multi-graph (as GraphML XML). Compared to list-like formats used in traditional archives, this approach offers a more intuitive representation of action dependencies and flows. This graph records all *actor-action-actee* behavior tuples described in the previous section, capturing all observed communication, DOM-modification, and script behaviors (i.e., calls to significant Web APIs and JS builtins) that occurred during the page's execution, detailed enough to allow the page's execution to be accurately replayed, step by step.

In conclusion, each of the three files included in each *.web* archive fulfills a different purpose to enable accurate, replicable, historical web measurement: the screenshot file records the presentation of the archived page, the HAR file records the resources loaded, and the GraphML file records the iterative behavior, over the page's entire lifetime.

4.4 Accuracy Improvements

This section demonstrates the improvements WebREC offers by comparing results from reproduced prior experiments with *.web* and common HTTP archives.

To do so, we revisit and reproduce three representative web measurement experiments from previous works, creating WARC, HAR, and *.web* datasets for comparison. We focus on three commonly measured resources: First, we analyze the accuracy of recording *JavaScript executions* by replicating experiments from Snyder et al. [217]. Then, we compare accuracy in recording *Document Object Model* content, a commonly measured resource, e.g., to analyze third-party roadblocks through inline event handlers for Content Security Policies [229]. Lastly, we measure how the *HTTP requests* differ and check how this helps the reproducibility of privacy research on third parties and trackers [67]. Before the experiments begin, we need a robust measurement pipeline.

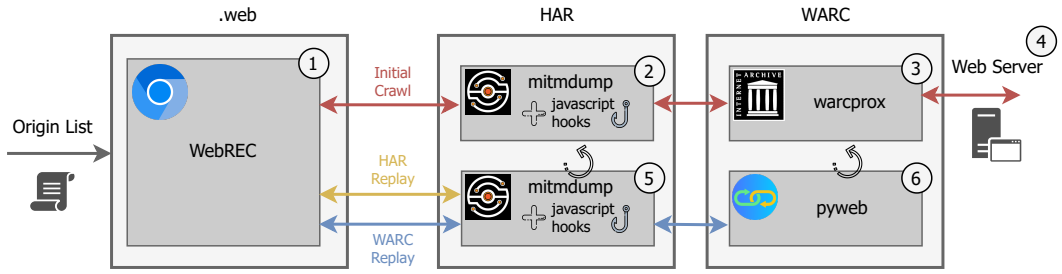


Figure 4.1: The pipeline used to create the initial dataset plus datasets for HAR replays and WARC replays

4.4.1 Dataset Construction

To perform meaningful analysis and comparison between different archiving formats, we need a dataset construction pipeline that meets several criteria. First, we need a *baseline* (BL) of recorded actions to compare all archives against. Otherwise, we cannot tell which archive is correct if two of them differ. Second, we require *consistent* recording of the same events during a single visit to account for non-determinism [198]. Third, the dataset needs to be *representative* for common web security and privacy measurement studies. To meet these criteria, we built a pipeline that creates a BL dataset from requests sent to the sites from the CrUX top 10k bucket, a highly accurate and literature-recommended website popularity list [201]. We proxy all requests through commonly used archiving tools to record the same requests (see Figure 4.1).

Our pipeline begins with an origins list (CrUX, December 2023). For each origin, it initiates a new instance of warcprox [103] (3), and mitmdump [154] (2), and then runs WebREC (1). We chose warcprox (from the Internet Archive) and mitmdump (part of mitmproxy) because they are well-maintained and tested tools. WebREC's requests are first proxied through mitmdump and from there through warcprox to the web server (ignoring TLS certificate errors). This setup ensures that all parts receive the same requests and responses, allowing us to record *.web*, HAR, and WARC files accurately.

With the *.web*, HAR, and WARC files, we can now compare requests and different file contents. However, it would not be possible to compare dynamic JS executions on a page for HAR and WARC files. To measure executions on archived content, prior work has made use of sites like the Internet Archive that replay recorded responses [197, 232, 138]. We apply the same method by using the replay functions of mitmdump for HAR files (5) and pywb [249] (recommended by warcprox) for WARC files (6), and then measure executions by hooking into monitored JS functions, keeping WebREC as unmodified as possible. The hooks are set via a prepended script element in front of every HTML response with the mitm proxy (2 & 5). Due to the error tolerance of browsers [S4], this script is executed even though it is positioned in front of any HTML tag. We note that this might lead to a difference in parsing due to the now entered quirks mode [150], a point that we discuss in Chapter 4.4.3. Additionally, the script also hooks into programmatically created iframes. To ensure that the script is always executed and not blocked by CSP, we use the Puppeteer option *setBypassCSP*

to deactivate CSP. This approach maintains consistency by using the same tools for the initial crawl and the replays, differing only in the response source; archive files instead of the live web.

For our baseline dataset, we use the JS (2) recorded with hooks during the crawl. Since WebREC is designed to avoid altering browser behavior, the JS hooks remain unaffected. After completing the pipeline for the CrUX top 10k, we have a solid dataset and baseline for analysis.

4.4.2 Ethical Considerations

Before we started our experiments, we carefully considered ethical implications it could have. In our study, we sent GET requests to the 10k most popular origins according to the CrUX list [201]. This is a research practice employed by numerous web studies before and is generally considered ethical in the security and privacy community. However, it is essential to acknowledge that ethical standards can evolve, and each project should be assessed on a case-by-case basis.

In the experiments in this chapter, we requested each origin of the top 10k only once per experiment with a standard GET request. Since a GET request is designed to receive data without making any modifications, we assume no negative impact for the end user. We also ensure no negative impact for the website operators as our minimal interaction does not contribute significantly to the common internet traffic or pose any more risk than typical internet background noise. While we, however, might potentially violate some websites' terms of service, as we did not verify whether automated site visits were explicitly allowed, we believe that the benefits to the research community in the form of our measurements and the demonstration of WebREC outweigh this concern. All our additional measurements are performed on our client side, i.e., our server, not influencing any third-party server. In conclusion, we are confident that our work did not cause harm to any individual or organization. Our research and measurements aim to contribute benefit to our research community by providing and influencing new ways how to conduct future web measurement studies.

4.4.3 Limitations

We also need to discuss the pipeline's limitations. As previously mentioned, adding a script tag at the beginning of each document causes the browser to enter quirks mode [150]. The quirks mode is an approach by browser vendors to ensure backwards compatibility and render even severely malformed HTML with their web browser. It causes the browser to follow other parsing algorithms and leads to slight differences in the parsed results compared to what typical users would see (e.g., CSS parsing). Yet, we are not aware of any relevant changes that would influence our results.

Furthermore, the pipeline misses executions inside of iframes added by the HTML parser if they contain a `srcdoc` property. Using a custom query with the HTTP Archive [97], we can see that only 10,534 (0.9%) of the 1,226,954 recorded iframes in their 1M dataset use `srcdoc`. While our pipeline could detect and rewrite these during the rewriting phase with mitmproxy, it could cause side effects introduced by HTML parsing libraries (e.g., by removing malformed elements) [S4]. Hence, we believe the

Appearances	*.web*		HAR		WARC	
Equal	39,603	93%	22,688	54%	22,645	54%
More	2,918	7%	1,934	5%	1,725	4%
Fewer	163	0.3%	17,557	37%	17,809	42%

Table 4.1: Appearances on the archived sites compared to 42,179 appearances on the baseline (BL)

impact of such mutations could be more significant to our results than being unable to hook into 0.9% of the iframes and thus decided against this approach.

Finally, given the multi-tier architecture, there is a small chance for a race condition in the resource recording. This causes a slight mismatch in recorded resources between archives and the *.web* format, leading to a very small difference in reported resources (details in Chapter 4.4.6.2).

4.4.4 JavaScript

Numerous web security and privacy measurements [219, 123, 218] analyze JS execution. In this section, we want to assess the differences, if any, in measured JS executions between reproducing experiments using a traditional archival format and our proposed *.web* format. To assess this, we replicate experiments from Snyder et al. [217], who investigated the prevalence of various JS APIs. The authors injected JS into each page to modify methods and properties to determine which API is used how often.

4.4.4.1 JavaScript Experiment

Our experiment focuses on APIs related to two JS standards: *CSS Object Model* (CSS-OM) [212] and *HTML Channel Messaging* (H-CM) [96]. Snyder et al.report that CSS-OM is often used and rarely blocked by privacy tools, whereas H-CM is very actively used but frequently blocked. This small scope allows us to highlight key differences between archiving techniques.

To count the API invocations after running our pipeline, we first analyze *.web*'s execution graphs for *js call* edges, which indicates API invocations. For the BL dataset and the replayed WARC and HAR responses, we count calls by analyzing the log files where the earlier mentioned JS hooks log every JS action on the page.

4.4.4.2 JavaScript Comparison

After running the pipeline on the CrUX 10k bucket, our dataset contained 8,479 *.web* files, each indicating that a crawl was successful. Next, we replayed the responses from HAR and WARC files. For 272 and 75 origins, respectively, replaying failed due to timeouts in reading the file (although set to a generous 60 seconds) or file interpretation errors. Due to overlapping issues, this leaves us with 8,150 successfully crawled and replayed origins for the remainder of this section.

For each origin, we counted the appearances of CSS-OM and H-CM API invocations. We define an *appearance* on an origin as the combination of the API name and the

Listing 5 DevTools log a CSSStyleDeclaration.

```
<svg id="s" width="100" height="100">
  <circle cx="50" cy="50" stroke="green" />
</svg>
<script>console.log(document.getElementById('s').style)</script>
```

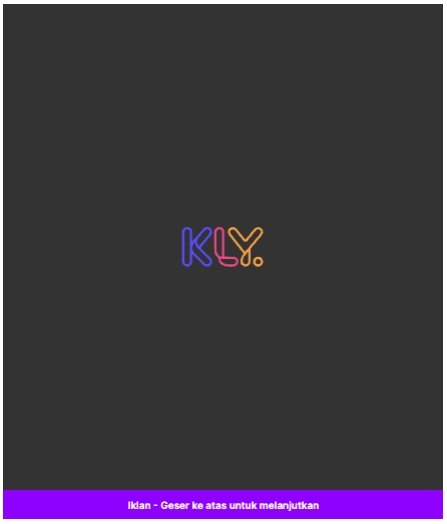
Figure 4.2: The DevTool logs show attributes of a CSSStyleDeclaration that are also logged with WebREC

number of times it was executed. Our BL dataset shows 15,775,881 API calls grouped to 42,179 appearances across origins. Table 4.1 shows that out of these 42,179 appearances the *.web* dataset aligns with 39,098 (93%) appearances on the same number of API calls, i.e., WebREC recorded the same number of invocations as the BL. For 2,918 (7%) of the appearances, *.web* captured more API calls. This can be attributed to WebREC's deeper hooking capabilities which, compared to the hooks implemented in JS, captures not only surface-level activities. For instance, the code in Listing 5 shows a snippet in which the SVG's *CSS Style Declaration* is logged, leading to the log output in Figure 4.2. The lookup for attributes that appear in the log, e.g., *accentColor* and *additiveSymbols*, are also recorded in *.web* while the JS hooks for the BL ignore it. Again, this difference is expected due to the underlying hooking differences. Moreover, only 163 (0.3%) appearances had fewer invocations in *.web* than in the BL, showing that researchers can achieve a nearly identical representation of API invocations as in the original visit, even in replication experiments conducted years later. Across all origins, the average difference in appearances between *.web* and BL is only 0.7%, highlighting the approach's accuracy.

After comparing the baseline to *.web*, we now focus on the comparison with replayed responses. Given that the underlying method to record the API calls is the same as for the BL, one would naturally expect results closer to the BL. The data shows that only 22,688 (54%) appearances in the HAR replays and 22,645 (54%) for WARC align with the BL, i.e., the overlap is significantly lower than for *.web* (93%). On the one hand, both HAR and WARC also have cases in which more API invocations are detected compared to the BL. We attribute this to the fact that replayed responses have no network delay when loading the resources, triggering some reoccurring API calls earlier.

On the other hand, the vast majority of appearance differences are because *fewer* invocations are detected in HAR and WARC; 17,557 (37%) and 17,809 (42%), respectively. Here, various of the reasons mentioned in Chapter 4.2.3 could be at play. For instance, our dataset reveals that sites that use Cloudflare (i.e., challenges.cloudflare.com) for bot protection tend to have fewer API calls for the replayed responses, likely due to the lack of real-time interactions with their servers, which are crucial for dynamic content exchange and behavior measurement. Another example is shown in Figure 4.3, a Google

(a) Ad from live website.



(b) Ad from archived website.

Figure 4.3: Google ads load dynamic content leading to a difference of JS executions between live and archived websites

ad element, first on the live visited website (4.3a) and then on the archived website (4.3b). While the archived version shows only a static logo, the live version loads an advertisement after a few seconds from the ad servers leading to various CSS-related calls. Especially for privacy related research, this difference could be relevant. And this difference in appearances is significant: While we observed an average difference across all origins of 0.7% between *.web* and BL, *we measured 13.9% for HAR and 13.3% for WARC replayed responses*. This highlights the drawback of replaying content rather than recording and inspecting the executions directly.

4.4.5 Document Object Model

Another typically analyzed component in measurement studies is the DOM of a webpage and its behavior, e.g., in case of DOM Clobbering [116], or with invalid HTML [S4]. To showcase how *.web* could be used for such experiments, we replicate a measurement from Steffens et al. [229], in which they analyzed to what extent third-party content on webpages complicates the security mechanisms Content Security Policy (CSP) and Subresource Integrity. In this section, we replicate one part of their hypothetical what-if experiment, in which the authors make the assumption that first-party developers want to deploy a CSP without using `unsafe-inline` and `unsafe-eval`. Their findings revealed that the vast majority of sites (86%) are unable to deploy CSP due to JS code, which adds inline event handlers to the DOM requiring `unsafe-inline`; we, therefore, focus in this experiment on JavaScript-added inline event handlers.

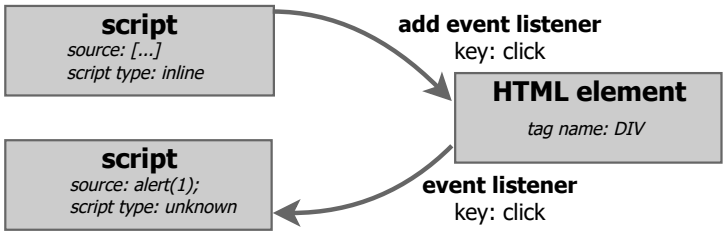


Figure 4.4: The representation of an EventListener as script node in the execution graph

4.4.5.1 DOM Experiment

To analyze inline event handlers, we run Steffens et al.'s open-sourced SMURF code in our pipeline to count all origins that use inline event handlers added to DOM through APIs like `document.write` and `innerHTML`. Using this code, we can simply count the origins in our log files and use it as a baseline for the comparison. We use the same method for the replayed responses from HAR and WARC files. Within *.web*, we query the graph for *event listener* edges (see Figure 4.4). If these point to the same JavaScript node that added the event listener (i.e., *add event listener* edge) or the script node does not have the type *unknown*, they are *programatically added*, i.e., do not involve string-to-code transformation. If, instead, they were added from the HTML parser or another script, they are actual inline event handlers, which would require `unsafe-inline`.

4.4.5.2 DOM Comparison

In this crawl, we successfully visited 8,523 origins with 311 HAR and 44 WARC (overlapping) issues replaying responses, similar to the previous experiment. Thus, this sections builds on 8,170 visits to analysis.

The results of the analysis are shown in Table 4.2. At first sight, the results appear extremely surprising: the original version of Steffens et al.'s code, SMURF [230], finds four times as many sites with inline event handlers (6,188) compared to WebREC (1,515). We, therefore, carefully analyzed SMURF's source code and found a number of impactful bugs. First, their code hooks into APIs which can be used to add elements to the DOM and subsequently checks if the added element contains inline event handlers. However, their code iterates over all the *properties* of an element instead of checking for *attributes*. This difference is critical, though: if an event handler is programmatically added through a function reference, this is a *property* of the element but not an *attribute*. In contrast, adding an inline event handler through `ele.setAttribute('onerror', '...')` (which requires string-to-code transformation) sets the attribute. A programmatically added event listener for an HTML node is a common way for many third parties to react to events while still allowing CSP to block inline scripts (a real-world example can be found in Appendix B.1). The SMURF code also does not properly capture event handlers which are nested inside a subtree of the DOM, leading to false negatives. Finally, SMURF assumes that any attribute starting with *on* is an event handler; as a result, our initial experiments showed that

Method	SMURF (bug)	SMURF (fix)	*.web*	WARC	HAR
Origins	6,188	1,644	1,515	1,472	1,368

Table 4.2: Origins for which we found inline event handlers

due to typos (e.g., *onclik*), HTML elements without meaningful event handlers were flagged as CSP blocking. Upon discovering these three issues, we contacted the original authors and fixed the code together with them to have a more meaningful baseline.

Given the updated code, we now investigate the feasibility of WARC, HAR, and *.web* to a study such as Steffens et al. [231]. We find that the replayed responses from WARC and HAR show fewer origins with inline event handlers, 1,472 and 1,368 respectively, than the initial baseline of 1,644 origins. This discrepancy is again due to dynamic elements, such as JS, not reloading accurately. These findings reinforce our argument that archive-with-metadata approaches are not ideal when the goal is to accurately reproduce web measurements.

For WebREC, there remains a difference from the baseline (1,644 vs. 1,515). Closer analysis of the findings showed that one issue relates to PageGraph, which was not able to handle event handlers inside SVG elements. We reported this limitation to the authors who addressed it in a newer version. Moreover, SMURF just logs any event handler regardless of whether it was added to the page's DOM and triggered (media elements are an exception) and irrespective of their origin, i.e., also reports those added in cross-origin iframes, which are not relevant for CSP. Hence, we believe the fixed implementation of SMURF to be in line with what WebREC can record.

4.4.5.3 Consequences

In order to conclusively refute the findings of Steffens et al., we would have to apply the fixed version of their code to the dataset from 2020. However, without a published dataset that is accurately reproducible like WebREC, we cannot make ultimate statements; even if the authors had recorded HAR or WARC files, it would be infeasible to fully replicate their findings. Nevertheless, we believe that inline event handlers played a less significant role for CSP blockage than indicated by Steffens et al.(assuming the Web has not radically changed since 2020). However, the authors also crawled subpages that could contain more event handlers and analyzed other aspects like inline scripts or eval calls, which we did not investigate. This makes us believe that the general takeaway of the paper remains valid.

In general, our findings show that the complexity of web measurements and JS hooking is prone to cause errors. Additionally, mistakenly collecting all *on** attributes or missing newly introduced event listeners cause problems in such traditional web experiments. As WebREC records all activities on a webpage, the list of event listeners is always accurate. Consequently, we believe researchers utilizing this technology would significantly improve their experiment pipeline by reducing their chance of implementing bugs.

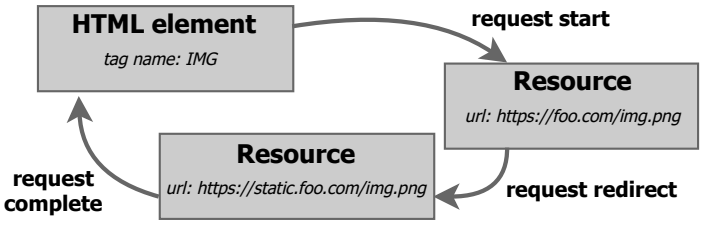


Figure 4.5: The representation of resource requests and redirects as a resource nodes in the execution graph

4.4.6 Requested Resources

Requests is another aspect frequently analyzed in web security and privacy studies for various reasons, for example, to support arguments about security headers [40], to identify privacy leakages [209], or to measure online advertisement and tracking behavior [67, 138]. For this section, we take the online tracking paper by Englehardt and Narayanan [67], a study in which the authors captured requests from one million websites and utilized the two tracking-protection lists Easylist and EasyPrivacy [66] to measure the prevalence of third-party requests and online trackers. Following their method, we can assess how well we can perform such an experiment using WebREC and show what additional benefits researchers could get from it.

4.4.6.1 Resources Experiment

Unlike the experiments before, researchers would typically not replay responses since all requests are stored in the HAR or WARC files already. Hence, we follow this method and only compare *.web* against HAR in this section without replaying responses. Additionally, we want to demonstrate how comprehensive WebREC's recorded page behavior is and use only this feature from *.web* to compare against HAR.

In these behavior recordings, we have two types of requests. Requests directed to load documents are tracked as URLs in the document node, i.e., the HTML pages themselves or any embedded document. Besides document requests, WebREC records resource requests and represents them via an edge to a resource node as demonstrated in Figure 4.5.

4.4.6.2 Correctness

Our results show that the requests in *.web*'s behavior records have great agreement with those recorded in HAR. In total, we count 778,500 requests in HAR after filtering out non-page-related requests like service workers or initial redirects, while we see 776,229 requests captured via WebREC for 8,544 successfully crawled origins. This makes a negligible difference of less than 0.3%.

Filtering the HAR file is required due to the very different ways these two file formats work. HAR is developed to record *all* requests that leave the browser, while WebREC captures every activity happening *on the page*. These do not include activities like

service workers, pre-flight requests, WebSocket upgrades, or CSP reports. Also, initial redirects are out of scope as they do not belong to the *page*.

We can attribute most requests made by the browser and filter them out from the HAR dataset by relying primarily on request headers. We attributed 12,342 pre-flight requests looking for the `Access-Control-Request-Method` header and 1,210 CSP report requests using `Sec-Fetch-Dest` headers set to `report`. WebREC behaviors do also not record the upgrade request for a WebSocket which we attribute via the `Upgrade: websocket` header, leading to 1,455 requests being filtered out. To attribute initial redirects from a document, e.g., *mobile.example.com* to *example.com*, we used the Sec-Fetch-Dest `document` and the `Location` response headers which gave us 2,819 filtered requests. Additionally, Chromium itself causes more than 40 requests per session, e.g., requests to the updater and similar, leading to 449,511 requests, which we attribute by looking for these URLs. Lastly, service workers also live in their own context and are not recorded in *.web* behaviors. While the initial loading of a service worker is detectable through the Sec-Fetch-Dest header, scripts included by the service worker do not carry such distinctive features. Therefore, we instead hooked into the serviceWorker API to disallow registration of any service workers.

Besides requests outside the page context, we also see requests that are internally made by the page but that never leave the browser. These never touch the HAR proxy but are recorded via WebREC. For example, some `Non-Authoritative Information` responses, like HSTS upgrades, are internally represented as the WebRequestAPI or `307 Internal Redirect`, likely Chromium internal behavior. Lastly, our manual investigation into differing requests shows a small race condition, a limitation in our pipeline: Sometimes, the browser sends a request just before we generate the *.web* output and tear down our pipeline. This means, a response might not return in time, resulting in an unrecorded HAR entry but an existing request in the behavior recording. Alternatively, the request might return and be recorded in HAR, but only after the *.web* is generated, leading to one more request in HAR. Due to these racy requests and the page-internal ones we could not to filter out, we measure a difference of less than 0.3% in the total number of requests.

This comparison shows that WebREC is well-suited for measuring activities directly *on the page*, while the proxy-generated HAR captures *all* requests – information often not needed for many studies, as the next section shows.

4.4.6.3 Replication

This section replicates a common third-party experiment to demonstrate that the page-context-only approach produces nearly the same outcomes as the more inflated HAR.

In the more than 700,000 requests we made to the top 10K websites, we see that 5,938 third parties are present on at least two first parties in the *.web* behavior records. At the same time, in HAR recordings, after filtering out the document redirects to avoid false positives (i.e., the false origin), we see 6,043 third-party requests. Same as for previous work, we can see that the majority of third parties are only present on a few sites as only 719 (*.web*) and 739 (HAR) third parties are present on more than 1% of all sites in our dataset. Next, we measured the prevalence of the third parties and compared the results to tracking-protection lists. The results for the top 20

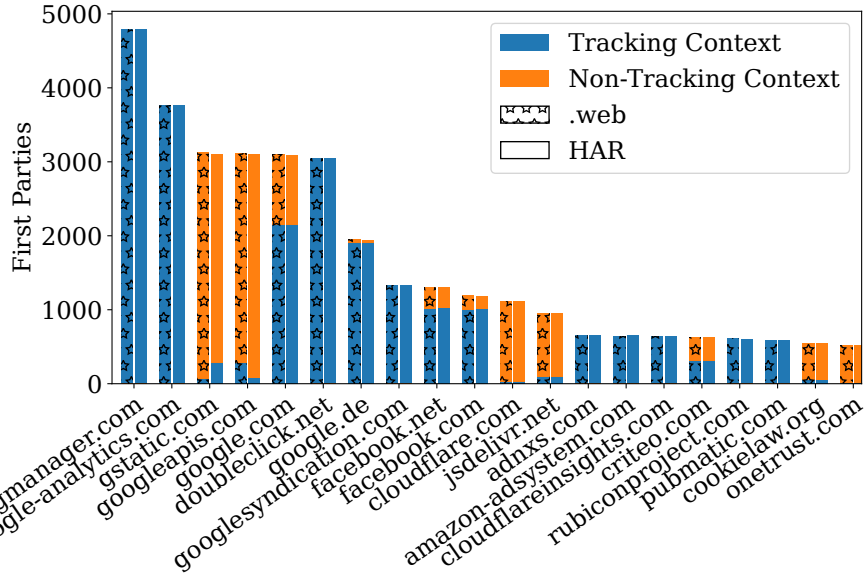


Figure 4.6: Top third parties on the top 10K sites via HAR and *.web*, divided in tracking and non-tracking context

third parties are demonstrated in Figure 4.6, showing the same observations as made by Englehardt and Narayanan [67]. As explained in their paper, the top third parties belong to only a small group of organizations. Indeed, the top 8 alone are associated with Google. Additionally, not all third parties are used in a tracking context, such as content delivery websites like *gstatic.com* or *jsdeliver.net*. The data confirms that whether measuring requests from the HAR recordings or utilizing activities captured with WebREC, the overall picture and takeaways remain consistent since the original study.

4.4.6.4 Attribution

In some experiments, simply identifying the request-target might not be enough. For example, if we want to understand the supply chain of third-party dependencies, we need to properly attribute the initiator of a request. However, it is non-trivial to accurately attribute requests to the corresponding source from which they were requested if the measurement is only based on recorded requests. Of course, indicators like the Referer header offer some insights, yet the Referrer-Policy could skew these signals. Consequently, a thorough understanding of third-party requests necessitates insights into on-page activities, activities recorded with WebREC.

We now demonstrate the additional insights that *.web* can provide compared to traditional HTTP archives. To that end, we analyze the requests sent towards the most prevalent third party, i.e., Google Tag Manager (GTM), which received a total of 12,564 requests from 4,642 origins.

Table 4.3 shows different variants through which the GTM is requested. Unsurprisingly, most requests come from script elements adding GTM. The most responsible

Context	Responsible Party	# Req.	# Origins
Script Element	*googletagmanager.com*	5,130	2,751
	first-party	4,342	3,503
	google-analytics.com	593	445
	adobedtm.com	77	40
	xhcdn.com	60	60
	...	1,534	1,169
IMG Element	*googletagmanager.com*	563	96
	halodoc.com	8	1
	googleoptimize.com	1	1
Fetch/XHR	*first-party*	157	1151
	sports.ru	2	2
	thesimsresource.com	2	1
	geekdo-static.om	1	1
Parser	*first-party*	92	74

Table 4.3: Third party requests to GTM grouped by context and responsible party who sent the request

party (causing 5,130 requests) is GTM itself, as the initial GTM script often includes additional scripts via script elements. It is followed by the script elements added by the first party, the expected way to add GTM in the first place. The long tail of third parties responsible for script elements (736 different third parties; not shown in the table), demonstrates the wide variety of entities websites use to load JS modules, including alternative tag managers like Adobe Dynamic Tag Management or own content delivery network such as *xhcdn.com*.

GTM is also used to load images (e.g., tracking pixels), mostly requested by GTM itself (563). Requests made through fetch or XMLHttpRequests APIs towards GTM are mainly caused by the first party (157). In addition, 92 requests were caused by parallel-loaded script elements (e.g., *defer*) within the page's original HTML (*Parser*).

This shows that WebREC's behavior recordings can be used for both to accurately measure requests and to further understand why and how these requests occurred by providing far more details than a basic request record does, which would have to rely on heuristics to determine this metadata.

4.5 Generality of WebREC

Previous sections showed that common web measurement methods lack reproducibility and sometimes correctness. We proposed WebREC as a general-purpose tool to overcome these issues, yet its generality remains in question.

To evaluate generality, we surveyed previous web measurement studies from top security and privacy conferences, building on a SoK paper by Stafeev and Pellegrino [225], who analyzed the use of crawlers in web measurement studies. They include USENIX Security, IEEE S&P, NDSS, ACM CCS, as well as the specialized venues WWW, PETS, and ACM IMC in their study, and collected all papers from 2010 to 2022 that contain the keywords `tranco`, `alexa` in a combination with `top` or `site`. From the resulting

1,057 papers, they removed all papers that did not employ automated crawling. The remaining papers build a list of 403 web crawling studies. Taking this list as a base, we focused on the time from 2020 to 2022 and thus analyzed all (137) papers to understand the individual uses of crawlers and whether or not WebREC could have been used. We only considered papers that measured data that are within the scope of what WebREC offers, i.e., they either analyze webpage JS executions, the page's content, or any HTTP requests made. This excludes studies using crawlers for other goals, such as network benchmarking purposes [151, 183] or side-channel attacks [238, 266] and narrowed our focus to 97 papers.

Our survey shows encouraging numbers, suggesting that 68 (70%) of the 97 papers could use WebREC for most of their data collection. This would not only standardize the process and reduce the chance for errors during collection but also save researchers resources and time. For example, Musch and Johns [166] investigated JS anti-debugging techniques in the wild by identifying code patterns and assessing performance changes, a tremendous development effort. Since these techniques are registered during page load, *.web* contains them, allowing this analysis without a custom crawler.

On the other side, studies unsuitable for WebREC include those measuring something outside the page context, e.g., service workers [115] and those focusing on very specific features, such as DNS over QUIC [130]. Additionally, since WebREC is based on Chromium, studies requiring other engines like Firefox are not catered for [167].

Even though 70% of papers could rely on WebREC instead of a custom crawler, the researchers would still require crawling, leading to avoidable resource usage and comparability issues. In fact, our investigation also shows that 47 (48%) papers could base their analysis solely on a *.web* archive instead, without the need for them to run a crawler. For example, Luo et al. [141] conducted an experiment labeling DOM elements as sensitive, subsequently checking what third-party scripts access them. Calzavara et al. [41] studied inconsistencies in security mechanisms comparing cookie policies and security headers across sites. Both DOMs and headers are collected by WebREC and could be analyzed in a *.web* archive.

To conclude, we believe that a significant percentage of web measurement studies (70%) could benefit from using WebREC over custom crawlers and that 48% could even be based solely on a *.web* archive. This highlights WebREC's potential to streamline measurement processes and to enhance reproducibility and correctness in our field.

4.6 Discussion

With the results of the previous section in mind, we now summarize the primary benefits of WebREC.

4.6.1 Reproducibility

WebREC addresses the problem of reproducibility by providing a comprehensive bundle of information about page requests and activities for researchers to use and share. Recorded at the core of the browser, this information offers a much more accurate representation of page behavior than existing archiving formats. For instance, for

JS API calls, WebREC shows an average difference of only 0.7% from the baseline, in contrast to a 13% difference with common HTTP archives. This discrepancy arises from dynamic behaviors like live server interactions, which can never be captured accurately by simply replaying responses. WebREC eliminates this need for replaying responses by storing all relevant activity information at runtime. Researchers can reproduce experiments by running the same scripts used in the original study on their dataset, ensuring reliable replication.

Going one step further than reproducing experiments, *.web* also allows researchers to delve deeper into the initially recorded data, offering a better understanding of various aspects of a historical experiment. We demonstrated for the requested resources experiment that we can not only replicate the key findings of the original paper but we also easily attributed third-party calls to different scripts, providing valuable insights into third-party deployment. This advanced capability can help researchers explain why replicated findings might differ over time, e.g., due to changes in the mixed content policy. We encourage researchers to provide their measured data as *.web* archives for future research to use.

4.6.2 Correctness

Since WebREC records behavior at the browser's core, the recorded behavior is correct eliminating the need for researchers to modify browsers or inject code into web pages.

Custom modifications can introduce issues if researchers are not fully aware of every browser-specific detail, leading to unintended consequences. For example, we showed, how a study [229] mistakenly attributed programmatically added event handlers as inline ones due to a mix-up between *properties* and *attributes*. WebREC avoids such problems by recording all page activities accurately at its core, allowing researchers to analyze *.web* without an in-depth understanding of the HTML or other specifications. In *.web*'s execution graph, the source of an event handler is clearly visible, reducing the risk of misattribution during analysis significantly.

Even if a potential bug occurs in the analysis phase, other researchers could easily reconstruct and correct it based on the provided *.web* bundles. That this is desirable shows the Steffens et al. [229] paper, for which we could not rerun their experiment with our corrected script due to the lack of an available dataset. Thus, we only run their script on our smaller dataset years later showing that over 50% of the visited sites in our data would have incorrectly attributed event handlers. How this would change the takeaway message of the original paper is unclear. We believe providing a format that allows other researchers to rerun and even correct experiments significantly improves the correctness of web measurement research.

4.6.3 Efficiency

WebREC improves the efficiency of web measurement research. The experiments in Chapter 4.4 demonstrate how *.web* supports various studies, eliminating the need for resource-intensive development of custom crawlers. Around 70% of papers we reviewed could use WebREC, highlighting the benefits of a standardized crawler for our community. On top of that, 48% of these papers conduct experiments that could rely solely

on *.web* archives without the need to live-crawl. We are confident that this will save resources, reduce web traffic, and prevent potential ethical pitfalls.

Such an archive would allow extensive future work, like longitudinal JS behavior analysis, currently not feasible with traditional web archives. It also allows easy prototyping of experiments without the need for additional crawls (which for example showed to be an resource issue on the IA in Chapter 3), thereby improving efficiency in web measurement research.

4.6.4 Storage Implications

Storing more information is expected to result in larger file sizes, requiring researchers to balance the new insights gained with storage demands. To address this, we added an optional compression step to our crawler and query tool. To estimate the expected storage requirements, we conducted another experiment comparing *.web* with the HAR files generated by a proxy. We ran our pipeline without any JS injections (which would otherwise inflate required space) on CrUX 10k, ending with 8,934 origins to compare.

For these origins, the average HAR file size was 13.6 MB, while *.web* files averaged 10.2 MB, broken down as 8.2 MB for HAR files, 1.5 MB for behavior records, and 0.4 MB for screenshots. The difference in HAR and *.web* HAR sizes is due to the fact that proxy-created HAR files include all browser requests, such as upgrade requests or updating dictionary files, whereas the *.web* HAR file only captures requests originating from the page context. This shows that WebREC, while providing more insights, keeps storage needs low.

4.6.5 Maintenance Overhead

WebREC is developed as a series of modifications to "stock" Chromium, along with external tools to interact with those modifications. Maintaining modifications in a rapidly changing code base like Chromium is a challenge, which can lead to quickly-abandoned prototypes, especially in research. To address this WebREC is designed as an ongoing solution, with design choices that dramatically reduce the maintenance overhead in keeping WebREC current.

Most of WebREC's complexity comes from modifications to Chromium and its subprojects (e.g., Blink, V8), which are mostly implemented in C++. The frequent changes to the code base make correctly maintaining modifications outside the project difficult. For example, patches that work correctly against one version of Chromium can need significant reworking for the next version, if Chromium developers have refactored code or added significant new functionality.

WebREC is implemented using strategies that simplify updates to new Chromium versions, and minimize amount of patching needed. First, most of WebREC's logic is implemented separately and independently of upstream Chromium code, kept in its own C++ namespace and data structures (instead of modifying the ones relied on by Chromium). Second, WebREC leverages many of Chromium's systems to capture information about a page's execution without needing to modify Chromium code. For example, Chromium's *CoreProbe* system allows code to receive notifications when certain page activities occur (e.g., new elements created). Third, Chromium's build system

and C++ classes include mechanisms to augment Chromium's classes in a more robust way than sub-classing. For example, the classes `Supplementable` and `Supplement` allow developers to add behavior to existing "Supplementable" classes, by implementing that new functionality in a "peer Supplement" class, without needing subclasses or changing upstream code, further making WebREC's behavior robust across Chromium changes. And fourth, when the previously discussed approaches are not possible, WebREC uses subclasses when possible, and modifies Chromium code through clever use of C/C++'s pre-processor (instead of through patches), both of which, while not ideal, end up being easier to carry across Chromium versions than patches. All this allows us to maintain WebREC with only 33 patches compared with over 12k lines to implement WebREC's browser-side functionality.

This design has been very successful in making WebREC easy to maintain across Chromium releases. WebREC has been kept current with every Chromium version since September 2022, (i.e., 40 Chromium versions) with little to no changes required each update. We estimate less than an hour of time has been needed to keep WebREC current for each Chromium release, many needing no additional effort at all.

4.7 Summary

This chapter addressed a significant issue in the reproducibility of the current web measurement landscape: dynamic content that is not correctly captured by existing archiving formats. That this dynamic content can easily lead to accuracy and reproducibility issues is shown by the presented experiments, which, for example, revealed a discrepancy of more than 13% in the appearance of JS API calls between responses replayed from common web archives and our baseline. Additionally, the experiments identified errors in a published paper, misattributing event handlers.

To tackle these issues and achieve reproducible and accurate web security and privacy measurements in the future, we then presented and evaluated WebREC. It bundles a web page's HTTP communication, an execution graph, and screenshots into a *.web* archive, proving more accurate in replicating page behavior than traditional archives. Our literature analysis shows that many existing measurement studies could have benefited from using it, offering a solution to current problems and improving future measurement research.

While, of course, more challenges exist, for example, encouraging researchers to actually apply our proposed solutions to make research more reproducible, this chapter concludes the part on reproducibility of web security and privacy measurements. We invite the community to create and share more datasets like ours to support future research. We believe that future work can greatly benefit from such datasets and archives, thereby enhancing reproducibility, transparency, and correctness in web security and privacy measurement research. After all, producing research that others can reproduce is itself an important aspect of responsible research, a topic we turn to in the next chapter.

Part II

Responsibility in Web Measurements

5

Legal and Ethical Challenges in Web Security and Privacy Measurements

Contribution of the Author

This chapter is based on "Where Are the Red Lines? Towards Ethical Server-Side Scans in Security and Privacy Research" [P2]. Florian Hantke, Sebastian Roth, Rafael Mrowczynski, Christine Utz, and Ben Stock conducted the study. Florian Hantke and Sebastian Roth developed the initial idea and refined the methodology together with Rafael Mrowczynski. Florian Hantke and Rafael Mrowczynski conducted the interviews, which were analyzed together with Christine Utz. All three designed and executed the survey, while Christine Utz analyzed the survey data. Ben Stock supervised the study. All authors contributed comments and text to individual sections.

5.1 Motivation

Confidence in science has decreased, in part, due to people's concern about the ethical acting of scientists. To avoid giving further grounds for these concerns, we need to better understand what the ethical challenges are in research – specifically, in this thesis, in web security and privacy measurement studies – and how we can address them to make our scientific practices more rigorous and responsible.

For web security and privacy measurement research, controversial ethical practices are particularly pronounced in studies targeting server-side issues. Unlike many client-side phenomena, server-side issues cannot easily be investigated in researcher-controlled environments. Studying them at scale requires interacting with unpredictable live systems operated by third parties, which introduces the risk of unintentionally harming remote systems. Such risks may lead to operator backlash, legal consequences, or problems in publishing research results.

Ethical considerations in S&P research have received increasing attention in the recent years, and several frameworks have been proposed to guide researchers [128, 182, 268, 50]. However, the boundaries of what is considered ethically acceptable or publishable are still discussed. For example, some researchers have attempted to study server-side vulnerabilities in real-world deployments, including work on Blind-XSS [120], SQL injections on orphaned web pages [186], HTTP desync issues [108], or ReDOS vulnerabilities [226]; some of these studies have anecdotally led to intense discussions within the research community and the program committee. Legal constraints further complicate the situation, as applicable regulations are often unclear and subject to interpretation [15], and even have led researchers to discontinue their research [76].

As a result of these uncertainties, server-side issues are conducted only at a limited scale, often within controlled environments of self-hosted open-source projects [68, 64, 148, 185]. While these approaches reduce ethical and legal risks, they provide only a small view of the Web, resulting in findings that provide only limited insights. This gap highlights the need for a clearer understanding of the legal and ethical boundaries of server-side measurement research.

In this chapter, we address this largely underexplored issue by investigating the feasibility of Server-Side Scanning (3S) from both ethical and legal standpoints. Our goal is to get a holistic understanding of the problem space and devise best practices

for future research based on 3S to improve the way we conduct server-side research to strengthen the confidence in our science. To gain broad perspectives on the risks and feasibility of 3S, we conducted 23 semi-structured interviews with German legal experts, members of Research Ethics Committees, and website and server operators to identify the boundaries within which such research might be permissible. To evaluate our findings at scale, we conducted an online survey with another 119 operators. Drawing on these insights, we suggest best practices and propose a pre-registration procedure to make 3S-based research more reliable and transparent.

In short, this chapter investigates the question, *How can we enable server-side scanning research within a framework that prevents harm for all stakeholders?* We investigate this via three sub-questions:

- RQ1: *How can server-side scanning-based research meet legal standards?* Researchers should not fear or risk legal repercussions. Yet, understanding the law, with all its nuances, can be challenging for non-experts. We therefore aim to understand the boundaries within which 3S research can be conducted legally.
- RQ2: *How can server-side scanning-based research meet ethical standards?* Research should follow ethical standards, which can be challenging in the complex web ecosystem where unintended side effects are not always easy to foresee. Hence, this work investigates concrete ethical challenges that researchers face.
- RQ3: *What problems do website and server operators see and how can they be minimized?* As operators are directly impacted by web research, their views are indispensable. We investigate what they consider harmful and permissible, aiming to address their concerns.

Our contribution comprises the following key points:

1. We present the first qualitative and quantitative study with 23 interviews and 119 survey responses that investigates diverse perspectives about 3S, (see Chapter 5.3.2), focusing on ethics committees of major security conferences and German jurists. While we use German law as a reference point in our study, the implications of our findings can be considered within a broader legal context.
2. We explore five concrete scenarios of how 3S is typically applied in web security research, along with expert and operator assessments. These can guide future work and help the community identify the general boundaries of 3S research (see Chapter 5.3.3).
3. We propose best practices on how to conduct 3S in security and privacy research (see Chapter 5.5.1) and suggest a pre-registration board for future work in the area (see Chapter 5.5.2).

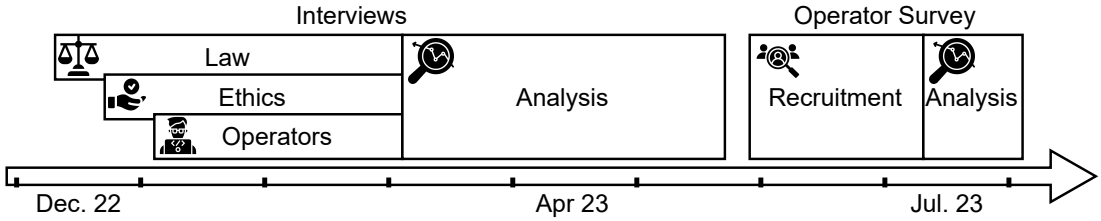


Figure 5.1: Study overview and timeline

5.2 Methods

Our study aims to provide guidance for server-side scanning (3S), focusing on use cases in web security and privacy. We investigate the boundaries of 3S from three perspectives, law (RQ1), ethics, represented by REC members (RQ2), and operators (RQ3). Our insights are grounded in interviews and an online survey.

To gain an in-depth understanding of these three perspectives, we first conducted semi-structured problem-centered interviews [256] with people from all three groups (see Figure 5.1). Our interviews with legal professionals and REC members had the character of expert interviews [26] aiming at the reconstruction of their specialist knowledge related to our RQs. Since legal regulations differ from one country to another, we narrowed the law-related scope of this study by focusing on German law. For the operators, semi-structured interviews provided insights into their mindsets and opinions on server-side research. Additionally, with far more operators existing than ethics and legal experts, we complement our findings with a quantitative survey study with operators to broaden the scope of opinions. As a key element to all interviews and the survey questionnaire, we use example 3S scenarios, presented in form of vignettes [196].

5.2.1 Vignettes of 3S Scenarios

Vignettes are "short stories about hypothetical characters in specified circumstances, to whose situation the interviewee is invited to respond" [73]. This instrument allowed us to elicit interviewees' considerations and judgments on specific 3S scenarios that are relevant to answer our RQs [6, 11]. For the survey, they let us approximate how well operators' opinions can be generalized on a larger scale.

We designed the vignettes to cover a range of typical 3S scenarios in web security research, which we identified by reviewing widely exploited real-world vulnerability classes known as Common Weakness Enumeration (CWE). Furthermore, the vignettes were meant to challenge participants' assumed mental models and therefore consider edge cases pointed out in ethics [14] and German legal [15] literature, e. g., the limits of access control, the importance of researchers' intent, and privacy law concerns. We also made sure to cover all dimensions of the CIA Triad. Each of the scenarios is identified by the name of its fictional protagonist (see Table 5.1). The full texts for all scenarios are provided in Appendix C.1. For the survey, we slightly modified the wordings to increase clarity; these versions are included in the full questionnaire in the supplementary material [92].

Table 5.1: Scenarios used in the interviews and survey

Scenario	Vulnerability	CWE Rank		CIA
Alice	SQL Injection	3	CWE-89	CA
Bob	Invalid HTTP Header	6	CWE-78	A
Charlie	Insecure Direct Object Reference	22	CWE-269	CI
Daisy	Stored XSS	2	CWE-79	CI
Eve	Path Traversal	8	CWE-22	C

5.2.2 Pre-Registration Proposal

Another integral part of both the problem-centered interviews and our survey questionnaire was our proposal for how to improve the status quo for 3S research. We propose a pre-registration process through a trusted third party (TTP) to provide a concrete framework within which 3S research can be carried out in a reliable and transparent way. The idea initially came up with insights from the first interviews and matured in internal discussion in our project team. In the later interviews and the survey, we briefly outlined the idea and presented our full proposal (see Chapter 5.5.2).

5.2.3 Problem-Centered Interviews

Our problem-centered interviews called for different recruitment strategies for each group of interviewees.

For legal experts, we compiled a list of professional roles we wanted to interview, e. g., legal scholars as well as legal practitioners like lawyers and prosecutors. We also included the perspective of academics, as (1) they are the most frequent contributors to *legal commentaries*, an important source of knowledge for all German legal professionals; (2) their *scholarly opinions* are regularly quoted in judicial decisions [117]; and (3) they also initiate debates that inspire legislative policies. Once the list was completed, we emailed legal experts who we identified via an online search and also followed further recommendations by our interviewees.

For operators, we followed a similar approach, creating a list of roles along with desired attributes (e. g., company in the security sector, small company). We found suitable operators online and emailed them, recruited via social media (Twitter and LinkedIn), and also used personal contacts.

Lastly, for the ethics perspective, we invited members from the RECs of USENIX Security and IEEE S&P, as they are the experts who assess the ethical validity of security research. To avoid conflicting too many potential REC members for our own submission, we invited one expert after another instead of sending out multiple invites at once.

For each of the three groups, we continued recruiting until we had filled all desired roles and attributes. If two people in one category had very diverse opinions, we recruited more people of this category until opinions saturated.

The guides we crafted for our interviews begin with general questions about web security, followed by more detailed questions tailored for each specific group. All five

3S scenarios and our improvement proposal were embedded in the guides; they are available as supplementary material [92].

We conducted one pilot interview per group to test the interview process and to fine-tune our guides. The interviewees were experts from our institutions. As this pre-study went smoothly, we kept this interviewing strategy.

Interviews took between 37–116 minutes and were conducted either in-person or via video conference between January and May 2023 (see Figure 5.1). With each new interview, minor adjustments were made to the guides, e. g., we incorporated new insights from earlier interviews. Most of the interviews were joined by two members of the research team. One person led the interview, while the other asked additional questions and handled the audio recording. We leveraged the Amberscript service for transcription. One interviewee did not consent to audio recording, so the (only) interviewer took notes during the interview and expanded them based on their memories afterwards.

5.2.4 Operator Survey

We conducted an online survey to validate our findings with a larger sample of operators. Like the interviews, the questionnaire was built around the five 3S scenarios and our pre-registration proposal for the future of 3S studies. To obtain more detailed insights into the boundaries of what types of 3S research operators are comfortable with, we leveraged ideas from the interviews to create two variants for each scenario: a less invasive one, which was shown to participants with a neutral or negative response to the base scenario, and a more invasive variant for neutral to positive respondents. The full questionnaire and all scenario texts are available in the supplementary material [92].

The questionnaire first asked about participants' background: if (Q1) and how many (Q2) servers they operated, their main role with regard to the operation of websites or servers (Q3), and the size (Q4) and location (Q5) of the largest organization for which they were currently operating them. We then defined 3S and asked for participants' general comfort level with researchers performing such scans on participants' servers, using five-point Likert scales (Q6). Next, we let participants assess the five 3S scenarios in the same way, presenting them in random order and asking for an assessment of the base scenarios (A1–E1); if applicable, one or both variants (A2–E2 / A3–E3); and an explanation of these assessments (A4–E4). The survey proceeded to present our proposal of pre-registration through a trusted third party. Participants were asked to rate the competence of different potential TTPs to assess 3S studies (Q7) and again to indicate their comfort level with such pre-approved scans (Q8). The section concluded with questions about factors that would be important to operators in such a pre-registration procedure (Q9) and if (Q10) and in what way (Q11a–b) it would change their opinion about scenario assessments. The survey ended with questions about previous security training (Q12–13), age (Q14), gender (Q15), and general feedback about 3S and our study (Q16). Participants could voluntarily provide an email address to receive the study results.

We pre-tested the survey in think-aloud sessions and test runs with operators recruited from the authors' social circles until no more comprehension issues arose. We

distributed the survey to operators at scale by emailing the generic email address `webmaster@DOMAIN` for websites on the Tranco top 1 million list [135] from June 21, 2023 (ID: X5XVN). We divided this list into 10 popularity bins of 100,000 domains and, once to twice a day, emailed batches of 5,000 domains, with 500 randomly drawn from each bin. This resulted in a total of 200,000 emails sent between June 23 and July 19, 2023, 159,009 bounced emails, and 291 survey accesses.

5.2.5 Data Analysis

Interviews. We implemented the procedure of qualitative content analysis [147, 206, 207], but in a relatively flexible manner. We derived our initial set of codes from our interview guides. These codes represented topics and sub-topics addressed by our questions.

After the codebook became saturated, a key concept of qualitative research [81, 236], three team members coded all interview transcripts in three consecutive stages. They had varied expertise with two being computer scientists, one also holding a German law degree, and the third being a social scientist specialized in sociology of law. Coding tasks were assigned such that every interview transcript was independently coded by at least two researchers. After each stage, we compared and discussed our codings until we reached an inter-subjective agreement. The final version of the codebook is included in supplementary material [92].

After finishing the coding process, we identified the key categories, i. e., the most important codes and used them to systematically look across different transcripts. In this way, we found commonalities as well as differences in what our interviewees had told us about 3S issues and our scenarios. The process of comparative data interpretation was facilitated by writing memos [81] on the main topics represented by our key categories. These memos served as the foundation for writing up our findings.

Survey. Overall, we registered 291 survey accesses. We removed 86 responses without consent, 85 partial responses that had not been submitted, and one response that had not answered a single question. This left us with a total of 119 final responses, which we analyzed using descriptive statistics. We did not conduct a formal qualitative analysis of the open-ended responses, as an inspection did not yield any new concepts beyond the interviews.

5.2.6 Research Ethics

Prior to recruiting participants for our interviews and the survey, we took extra care to minimize any potential harm or privacy concerns. We obtained prior approval from both our institution's Ethical Review Board and data protection officer. Both interviewees and survey participants were briefed about the goal of the study before their participation and asked for their consent. They also had the option to withdraw from the study at any time. One interviewee was not comfortable with being recorded, so we respected that wish and resorted to taking notes. The operator survey was distributed to popular websites by contacting only the generic email address `webmaster@DOMAIN`. While this could be considered as unsolicited bulk mail, our use of email generics meant for general public inquiries was found by a data protection authority to be less invasive

than more targeted approaches [247]. Each address was only emailed once, without any reminders or confirmations. Email addresses voluntarily provided by survey participants to learn about the study results were only contacted once to send out a preprint of original study and deleted afterwards.

5.2.7 Limitations

The qualitative part of our study has general limitations that characterize this type of research: We can identify different arguments with interviews, but we cannot make generalizing inferences about their distribution within the entire population of legal experts or REC members. For example, we were able to identify arguments rooted in different understandings of research ethics, but we cannot indicate which direction dominates among REC members.

Our inquiry into legal assessments of 3S research is also limited in terms of geographic scope. We exclusively interviewed experts on German law, due to practical reasons. Despite all transnationalization trends of recent decades [24], law remains a fairly national phenomenon and most legal professionals operate within their respective national legal systems [1, 2]. A comprehensive assessment of relevant laws would require a vast multi-national research team and massive resources, which we do not have. Thus, we decided to focus on Germany as an example. We hope that this work inspires researchers globally to conduct similar studies in their countries, painting a broader picture piece-by-piece.

As for web operators, we combined semi-structured interviews with a survey. This mixed-methods approach aimed to overcome the limitation in the qualitative part. However, our survey also has its own methodological limitations. Despite the random selection of potential respondents, self-selection bias necessarily occurs. In particular, we suppose that security-aware operators were more likely to respond.

5.3 Results

The insights obtained in 23 semi-structured interviews and 119 survey responses paint a consistent picture of the legal and ethical challenges of server-side scans and the potential problems for operators. In this section, we summarize the participants' views, emphasizing that these should not be misconstrued as legal counsel. First, we describe our participant samples and the three interviewed groups' general stance towards 3S. Afterwards, we go into more detail, focusing on the five 3S scenarios selected to represent common server-side issues in web security. We conclude with participants' suggestions for improvement and opinions on pre-registration.

5.3.1 Participant Samples

Interviews. As shown in Table 5.2, we interviewed a total of 9 legal experts, 10 operators of various websites, and 5 members of conferences' RECs. Considering the small number of the REC members in general, we opted to omit the country's name to prevent the de-anonymization of any participants. We also spoke to other people who did not fit our three designated groups, e. g., politicians. While not included

Table 5.2: Interviewee demographics and background

	ID	Role	Gender	Country
Law	1-L	Law professor (criminal law)	M	DE
	2-L	Law professor (privacy law)	M	DE
	3-L	Law professor (criminal law)	M	DE
	4-L	Law professor (privacy law)	M	DE
	5-La	Law professor (criminal law)	M	DE
	5-Lb	Legal research assistant	F	DE
	6-L	Legal practitioner	M	DE
	7-L	Legal practitioner	M	DE
	8-L	Legal practitioner	M	DE
Operators	9-O	CISO of a large company	M	DE
	10-O	Hobbyist and self-hoster	M	DE
	11-O	Operator for web agency	M	DE
	12-O	Pentester and web operator	M	DE
	13-O	Owner of multiple web shops	M	DE
	14-O	CISO in the public sector	M	DE
	15-O	CTO of a small startup	M	DE
	16-O	CEO of a web agency	M	DE
	17-O	CTO of an international company	M	UK
	18-O	Pentester and self-hoster	M	DE
Ethics	20-E	Ethics Committee Member	M	-
	21-E	Ethics Committee Member	M	-
	22-E	Ethics Committee Member	M	-
	23-E	Ethics Committee Member	M	-
	24-E	Ethics Committee Member	M	-

in our analysis, they still contributed to our overall understanding. Interview 5 was conducted with two participants simultaneously, resulting in 23 interviews but a total of 24 participants.

Despite our efforts to achieve a diverse participant pool by sending out emails to a broad spectrum of people and also advertising on Twitter (>40K impressions) and LinkedIn (∼2K), all but one interviewee were male – a reflection of the unfortunate gender disparity in these fields. Interviewees' ages ranged from mid-twenties to retirement, and they were based in Germany, the UK, the US, and Switzerland.

Participants were diverse in terms of role, professional background, and seniority. Among the legal experts, 6 hailed from academia, ranging from research assistants to senior professors. The practitioners worked as lawyers and prosecutors. The REC members also ranged from junior to senior professors. The interviewed operators held a variety of positions and had a diverse understanding of security. The group included self-hosters, CTOs, CISOs, penetration testers, web developers, and one owner of multiple online shops. Overall, our groups of interviewees covered all the roles we had planned to interview.

Survey. After data cleaning as described in Chapter 5.2.5 we were left with 119 valid survey responses. Participants here also predominantly (84.9 %) identified as male and had a mean age of 43.0 years (std 9.8, min 20, med 43, max 72). All but one participant reported to operate a web server or site; within the last three years, they had operated a mean of 343.0 servers (std 1809.5, min 1, med 15, max 16,000), with most being responsible for 2–50 servers. Most (92.4 %) reported to have received prior security training, most frequently via self-teaching (89.1 %) or "learning by doing" (77.3 %).

Formal education was less common, led by courses at university or school (48.7 %). Appendix C.2 contains detailed statistics on survey participants' demographics and background in server operation.

5.3.2 General Assessment of Server-Side Scans

In the first part of each interview, we introduced the interviewees to the topic of 3S and asked what they thought in general about researchers conducting such scans. Depending on their field, this already provided evidence of diverse perspectives and the breadth of affected subject areas. The following sections describe interviewees' general assessments of 3S, while Chapter 5.3.3 provides more detailed insights discussing specific 3S scenarios.

5.3.2.1 Legal Experts

Our interviews with legal experts indicated that topics like 3S-based studies, searching for server vulnerabilities, and ethically hacking activities constituted a *"niche topic"* in German juridical debates (3-L). Currently, this kind of research can only be done in a *"legal gray zone"* (7-L) – a fact that implies serious legal risks for those who conduct 3S-based studies (7-L; 8-L).

On the one hand, a systematic search for vulnerabilities on remote computer systems, especially if performed at scale, can constitute an act of crime under present-day legislation (5-La). The only legally sound way to perform 3S-based security research would require explicit consent by every single operator whose server is subjected to a scan (7-L). However, 7-L underlines this is infeasible for 3S studies.

On the other hand, as of August 2023, there are no public court rulings on 3S-related cases in Germany (5-La)[1]. As a prosecutor interviewee emphasized, no *"white-hat hacker"* has ever been sentenced by a German criminal court, as all such rare cases had been already closed before trial, either because of a lack of "public interest" in prosecution or because of "minor guilt" – both are discretionary prerogatives often available to prosecutors (6-L). This fact, however, has ambiguous implications for 3S researchers: A conviction appears unlikely but there is no guarantee (5-La), and a *"mean-spirited"* prosecutor could still open an investigation, which by itself is already stressful for suspects (e. g., confiscation of their hardware), even if the case is later closed before trial (2-L).

The biggest legal risks and uncertainties for 3S researchers lurk, however, in the civil law domain. If a scan causes harm on remote systems, affected operators can sue for compensation (6-L). Harm can include different things: a need to pay for IT specialists to fix a server, a direct loss of revenue due to the unavailability of an operator's Internet presence, as well as a loss of customers due to reputation damage (6-L). The distinction between intent and negligence, which is vital in criminal cases, or notions of public interest or benevolent motivations are barely relevant in civil law cases (6-L).

[1]This changed in 2025: German courts have consistently upheld convictions in a prominent case in which an IT consultant was found criminally liable for accessing a publicly available database password, despite subsequently reporting the issue responsibly. [204]

5.3.2.2 REC Members

Besides the 3S topic, we asked the interviewed REC members about the process of ethical review. They are selected to serve on the REC based on their research experience. Although they usually do not have systematic academic training in practical philosophy, some of them do explicitly refer to main ethical traditions, as indicated by one of our interviewees and in literature [128]. Other REC members we interviewed perform their tasks based on implicit understandings of ethical norms without explicitly linking them to broader philosophical debates.

According to our data, there are two main ethical traditions influencing the REC members. The first is *utilitarianism* [23, 153] (a sub-form of *consequentialism*) [128]. It focuses on relations between benefits and harms as consequences of human actions, e. g., a 3S-based study. The second philosophical tradition is *deontology* that emphasizes the *categorical* moral obligation to respect the fundamental personal rights [114], e. g., privacy. "[D]ifferent ethical frameworks can lead to different conclusions about right and wrong" when discussing ethical challenges posed by specific research projects [128].

Other ethical frameworks were also mentioned: *"discourse ethics"* and *"ethical colonialism"* (22-E). Proponents of the discourse ethics [9, 86] argue that an adequate ethical assessment can only emerge as a result of a free and open-ended discussion between all stakeholders affected by that action. The term *ethical colonialism* is understood by one of our interviewees as a critique of a situation in which *"people who are in a position of power [are] imposing their worldview on other people"* (22-E). This notion seems to be closely related to what is also denoted as *ethical imperialism* [214] or *moral imperialism* [109].

Our data shows that utilitarian and deontological arguments are frequently made by all interviewed REC members, even if they do not explicitly reference any philosophical sources. Usually, reasoning about ethical problems posed by our scenarios is rooted in both major ethical traditions.

Discussing the 3S research also highlighted the ethical challenges for researchers. Uncertainty about a possible impact of research-related 3S activities on remote systems was mentioned as the key problem (20-E, 22-E). This results in an ethical dilemma: *"Is the morally right decision to not scan because not scanning would then mean that the researchers are not involved in the actual crashes of any machines, or is the morally right decision to scan, thereby identifying vulnerable machines and encouraging them to fix and thereby helping other users?"* (22-E).

There are also other general ethical problems identified by one interviewee: (1) absorbing server resources for processing of numerous requests; (2) binding human resources by generating many log events which can alarm security personnel; (3) researchers lacking resources for a large number of vulnerability disclosures (24-E). Another interviewee (23-E) named three types of 3S activity that constituted absolute red lines from an ethical point of view: (1) *"fuzzing random stuff"*; (2) intentionally damaging someone's business operations; (3) extracting any kind of personal information.

Still, our interviews revealed that *"the red lines are not clear"* (24-E) and *discourse ethics* played an implicit role in all interviews with REC members. Researchers performing 3S should report in their papers any incident caused by their research activities (e. g., server crashes or privacy breaches), because only under this precondition can the

REC become aware of harmful side effects of a given study and adequately assess its ethical implications. Researchers should also justify their selected research design by demonstrating that it is best suited to keep up high ethical standards (24-E).

5.3.2.3 Operators

Overall, the interviewed operators mostly exhibited a positive attitude towards 3S conducted by academic researchers. Six interviewees explicitly mentioned that they saw it as their own responsibility to secure their services once they go public on the Web. They acknowledge that such scans are part of Internet reality and operators should be well prepared, because *"[a]t the end of the day, the bad guys do it."* (18-O).

Most operators understood the purpose and necessity of 3S activity, as it contributes to securing the Web in the future. Still, operators voiced concerns about potential harm. For example 16-O said: *"I think [3S] is absolutely fine. Our system must also be able to withstand this to a certain extent. If you basically say, '[3S] is legal,' I would have a problem with that."*

Issues related to infrastructure were noted in all interviews: *"[Y]ou never know what kind of business workflow you're going to trigger behind the scene."* (17-O). This lack of knowledge could lead to unintended consequences such as server crashes. Scans should make as few requests as possible to avoid accidental server overloads or denial of service. A red line would be crossed if researchers risked the system's availability on purpose. *"Anything that goes in the direction of denial of service, performance degradation and the like is problematic"* (11-O).

Another core aspect is the effects related to the end-user: *"[S]ecurity research should not interfere with websites' users"* (13-O). One frequently mentioned red line is the leakage or manipulation of private user information. Researchers should not actively search for personal data. If such data is discovered, they should handle it confidentially.

Operators' third main concern is harm to the organization. This could be reputation damage, when vulnerabilities are leaked to the public (9-O; 13-O), thus, researchers should handle vulnerability information confidentially (9-O). 3S could also cause financial losses for the organization. For example, as cloud services charge by traffic, large scans can cost organizations money (15-O). Many organizations also hire commercial pay-as-you-go CERTs to monitor their infrastructure, which means alarms appearing from 3S could lead to unnecessary costs (15-O).

Seeing the potential damage, operators also mentioned how they would react to 3S activity. Some highlighted their right to block IP addresses; others pointed out they had already implemented mechanisms to automatically block aggressive scans. Survey respondent 1-OS noted, *"I would actively try to stop it from my side. I believe you have the right to try and I have the right to try to not allow it."*

Besides technical reactions to scans, operators mentioned that they would seek communication with the scanning party if necessary, e. g., when noticing aggressive scans or harm done (12-O; 13-O). Although the filing of criminal complaints was mentioned, it appears to be less common and is usually reserved for events of significant harm or threats to reputation. If operators see that researchers are behind a scan, e. g., indicated by a header, direct communication is the preferable first response, with criminal complaints seen as the last resort. Nevertheless, two operators mentioned that in case

of significant harm, they would be obligated to file criminal complaints regardless of the scanner's intention as their company's legal team would ask for it (14-O; 15-O).

Although the interviewed operators see the negative consequences, we already pointed out that the majority of operators in our study are positive about such scans, if conducted by academic researchers. They would rather accept the potential harm from researcher-conducted scans, knowing that it would be responsibly disclosed, than risk damage caused by criminal actors. Not only do the interviews indicate these opinions, but also the survey data in Q6. As shown in Figure 5.2 and Appendix C.4, more than half of the participants (36.1 % comfortable, 21.8 % somewhat comfortable) were positive about researchers conducting server-side scans, while 9.2 % were neutral and about a third expressed a negative stance (13.4 % somewhat uncomfortable, 19.3 % uncomfortable).

Key Takeaways

Legal assessments of 3S remain problematic due to a lack of criminal precedents. Even more critical is the risk of civil lawsuits. REC members stress the discussion of researchers' ethical decisions and transparency in publications, wishing for a well-argued balance between potential harm and benefits. The operators group is the most positive group in our study when it comes to 3S, as they also benefit from it in the form of vulnerability disclosure. They view legal actions as a last resort in cases of significant harm or if required by the company's legal team.

5.3.3 3S Scenarios

At opportune moments during the interviews, we asked the interviewees about their opinions on the five ealier mentioned 3S scenarios. This allowed us to follow their trains of thought in assessing the concrete risks and benefits of 3S. Before reading the assessments, we would like to encourage readers to first think about each scenario and form their own opinion.

5.3.3.1 Alice – SQL Injection

In the first case study, we asked participants about the use of the SQL `sleep` function to test for the presence of SQL injection vulnerabilities.

The majority of legal experts in our study agreed that these types of scans had a low criticality from a legal standpoint. This is because Alice did not read any sensitive information from the server, which neither constituted the offenses of data espionage or phishing nor violated privacy legislation. Moreover, the potential for harm to the infrastructure is low, reducing the risk of a civil lawsuit. Still, several interviewees noted that a strict interpretation of the criminal code could lead to Alice's action being considered an unauthorized manipulation of data, as one could *"deliberately delay the response now and activate some particular mode in the database [...]"* (3-L).

While the legal experts attested the SQLi approach a low risk level, most concerns raised by the REC members in our study referred to legal issues: *"Alice should also make sure that she obeys all the laws in the country"* (20-E). They were also concerned

that Alice had no control over server reactions. While most servers would likely just execute the sleep command, *"how can she make sure that the server does not crash or maybe misbehave"* (20-E). Consequently, they concluded that the REC would urge authors to consider other methods to address the research question and ask: *"[D]o you really need to do this to get your answer or are there safer things to do?"* (21-E). Despite the raised concerns, two of the REC members considered the sleep function *"a good way of minimizing the risks"* (24-E). They would accept such a paper and rather advise authors to disclose any incidents.

The interviewed operators mentioned arguments similar to those of the REC members. Three of them pointed out that researchers could neither predict how the infrastructure would respond to a request nor the implications on log files or caching. For instance, 9-O warned that if the payload was more than a sleep call, *"internal attackers [...] are suddenly able to get stuff [from the log]."* Operator 11-O expressed concerns about an even less invasive variant using `SELECT 1+1` instead of `sleep`. It potentially disrupted caches and, consequently, *"some subpage is no longer the title of the post [...] but 2"* (11-O), impacting end users. Regarding the initial sleep payload, another concern was that Alice did not know how time-critical a service was (17-O). In some real-time services like stock trading, a one-second delay could already cause significant financial damage. However, 16-O challenged this argument, stating that the natural latency on the Internet could often exceed one second. Despite these concerns, the common consensus was that a short sleep was acceptable, as it did not significantly impact end users. The `1+1` variant was even better received by the majority of operators, yet a red line would be crossed as soon as personal information was leaked, with some already considering knowledge of the database structure as critical.

These views are further supported by our survey data, as shown in Figure 5.2 and Table C.8 in Appendix C.4: 69.7 % of the surveyed operators reported being (somewhat) comfortable with the initial sleep scenario (question A1), while 22.7 % were (somewhat) uncomfortable with it. The less invasive `1+1` case (A3) yielded a (somewhat) comfortable assessment by an additional 4.2 %, with 18.5 % still being (somewhat) uncomfortable. The more invasive variant in which Alice reads the database structure (A2) still led to some degree of comfort with 37.8 % of respondents and 29.6 % being (somewhat) uncomfortable with it.

5.3.3.2 Bob – Invalid HTTP Request

In this scenario, Bob sends invalid HTTP requests, accidentally crashing a server.

The interviewees from law all agreed that it came down to Bob's intent and potential negligence. For the judge in a criminal case, it was important to understand whether Bob knew and expected what would happen or not: *"[I]t depends very much on [...] the probability [of a crash]"* (3-L). If Bob had known that the server might crash and still sent an invalid request, or if this even was his intention, it could be argued that he committed computer sabotage. Regardless of the legal status of the act of sending such a request, civil law enables operators to demand compensation for caused harm or file an injunction to stop any 3S activity against their servers if Bob acted with intent or negligence.

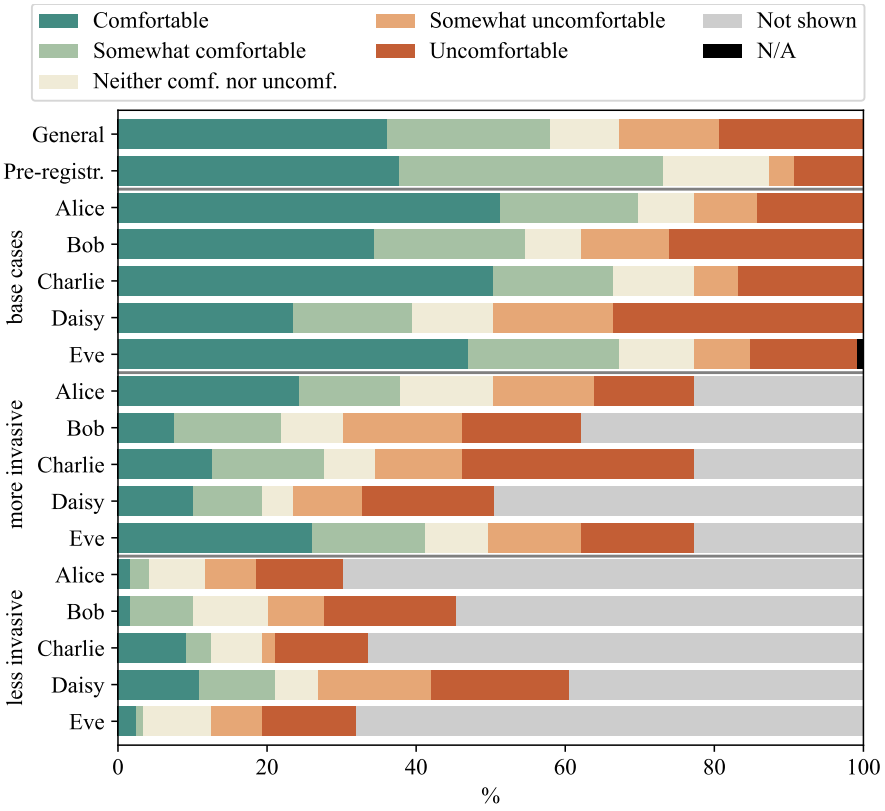


Figure 5.2: Surveyed operators' level of comfort with 3S in general (Q6, Q8) and 3S scenarios (A1–E3). The less and more invasive scenarios were only shown to participants who already were uncomfortable / comfortable with the base scenario, respectively (see Chapter 5.2.4).

The interviewed REC members expressed similar views and highlighted the importance of due diligence to demonstrate good intent. *"[M]any reviewers would first ask the question [...] what diligence did Bob do to mitigate harms beforehand?"* (22-E). Interestingly, they *all* asked this question, concluding that *"[i]f [Bob]'s not aware of this problem at the beginning, then it would be ethical"* (24-E), as his intent was not to crash a server. However, as soon as Bob's monitoring indicated a crash, he should stop scanning, disclose the issue to the operators, and mention the incident in the paper such that others would not make such mistakes.

In the same vein, the interviewed operators agreed that sending invalid HTTP requests would be fine. In the real world, something similar could also happen due to bugs in end-user software like browsers (9-O). Thus, operators should be prepared for such behavior and researchers should be allowed to test for it. *"I understand that's how research works, you can't make everything 100 % safe"* (15-O). The operators in our study further mentioned that they could also learn from such crashes and prevent them in the future. For example, Survey participant 97-OS wrote, *"If it is a crash that is reliably replicated [...], I care more about knowing about it than the service being temporarily down."* However, most operators would only be fine with a crash if it

occurred just once. Thereafter, they would expect researchers to stop.

The survey responses paint a similar picture, as 54.6 % of participants would be (somewhat) comfortable with Bob's scan in the base scenario (see Figure 5.2 and Table C.8). However, 37.8 % expressed (some) discomfort, probably because this scenario involved a crash. The less invasive variant (B3) only convinced another 10.1 % of participants and still was viewed negatively by 25.2 % of participants. On the other hand, the more invasive scenario (B2) still elicited a (somewhat) comfortable response in 21.8 % of surveyed operators, with 31.9 % being (somewhat) uncomfortable.

5.3.3.3 Charlie – Insecure Direct Object Reference

In this scenario, Charlie first changes a GET request that contains his user ID to instead use the ID of another user. This checks if it is possible to receive information from other users. In the more invasive variant, Charlie changes the ID in a POST request to check if he can modify other users' data.

For the GET request, all interviewed legal experts agreed that from a criminal law perspective, the possibility of charges for data espionage depended on whether an ID could be considered access protection. *"The question is, is it already enough as access security that a standard Internet user [...] cannot do that?"* (3-L). Six concluded that a predictable ID by itself could not be considered sufficient access protection but pointed out that other experts might disagree. Thus, Charlie's first scan would *"probably [go] unpunished, because you cannot circumvent any access protection"* (6-L).

However, the POST request that modified data would unquestionably be illegal, even without bypassing any access protection. As 1-L exemplified: *"If I were to leave my laptop here unencrypted and you [...] made an A into a B, that would already be a [illegal] data manipulation, regardless of whether there is password protection or not."*

Beyond criminal law, for both scans, the interviewees referred to privacy laws and their complexity. Some regulations, such as Article 89 GDPR [70], hold exceptions for researchers, meaning that with sufficient justification the processing of users' personal data for research is legal.

The REC members also considered the GET request to be less harmful than the POST idea. Two interviewees would accept such a method in a paper without discussion, *"[h]e just needs to make sure that the handling of [...] the sensitive information [...] is also okay and then also describe it in the paper."* (20-E). Yet, they mentioned that this would likely lead to discussion in the committee. For the modifying request, all members agreed that it would cross a line. However, they all would consider alternative research designs, asking: *"Is there a reason Charlie didn't create two accounts that he then tries to change between the two?"* (22-E). All interviewed REC members stated they would be fine with such research if Charlie instead used a dummy account to evaluate his experiments against. This would not harm any individual users and, thus, would ethically be acceptable. Nevertheless, participant 22-E brought up the question of potential terms of service violations. Asked about this, legal expert 5-La expressed a similar view, acknowledging that while the use of a dummy account might not constitute a criminal offense, it could potentially violate the service's terms and conditions.

Asked about Charlie's experiments, no operator mentioned terms of service. Three of them were fine even with the data modification, and five with the GET request, as every operator should check for such flaws. Three operators mentioned privacy concerns that would speak against the scans, because *"as soon as I notice that users' data is being leaked, I have to run to the data protection authority [...]. I don't feel like doing that."* (10-O). Operators also supported the dummy account solution, especially if they saw a connection between two accounts, e. g., similar names.

Survey data (Figure 5.2) confirms the interview sentiments, with 66.4 % of participants being (somewhat) comfortable with Charlie's GET request (C1) and 22.7 % expressing (some) level of discomfort. As expected, comfort levels were lower for the modifying POST request (C2): 42.9 % of operators expressed (some) discomfort with this, but still 27.7 % felt (somewhat) comfortable. By contrast, Charlie only accessing his own dummy profiles (C3) yielded an additional positive response of 12.6 % but still caused some level of discomfort with 14.3 % of surveyed operators.

5.3.3.4 Daisy – Stored XSS

Daisy enters an XSS payload into an input field on a website, subsequently storing it on the server. This payload potentially reaches all site users and, upon execution, sends a ping back to Daisy's server.

The legal experts in our study viewed this case as problematic. All but two of them pointed out that Daisy would be infringing on privacy laws by collecting users' IP addresses. *"The IP address, whether static or dynamic, is personal data"* (4-L). While the Article 89 GDPR does make exceptions for researchers, any advancement of scientific research must be weighed against the individual's right to privacy. The experts were not convinced of the benefits of Daisy's research. From a criminal law perspective, all but one expert considered uploading data explicitly deemed to be an attack payload an illegal manipulation of the server-side data: *"[T]his code is stored somehow [...], individual bits and bytes are actually changed without the user's consent"* (1-L). As Daisy's actions did not only target a server but potentially all of its users, she may be subject to prosecution for multiple offenses against the server and each individual client.

The interviewed REC members assessed the scenario similarly, stating that researchers should not store potentially harmful code on a web server. They expressed concern that once the code was stored, it left a lasting sign of an attack on the server. Like the jurists, they found it very concerning to execute code on multiple clients. Only one interviewee (24-E) said they would be fine with Daisy's experiment if the method was the only viable means to answer the research question and potential benefits outweighed the potential harm. This participant would require Daisy to provide more information explaining that her research was ethical and her method represented the least risky approach, as *"ultimately, it's the responsibility of the authors to demonstrate that their research is ethical"* (24-E). If Daisy could revise her design to ensure she does not inadvertently attack unknowing users, most REC members we interviewed would not reject the paper on ethical grounds. One option could be a check added to the payload that ensured only browser(s) controlled by Daisy send a ping back. Other

browsers would still execute the confirmation code, but the harm would be acceptable for most of the interviewed REC members.

Operators in our study were also mostly comfortable with the proposed alternative research design. *"It is okay. The way you are choosing the payload shows your intention, I think"* (17-O). In the base scenario, operators expressed discomfort with a malicious payload being stored on their servers, similar to the REC members; yet, they were more troubled by the privacy implications pointed out by the legal experts. This sentiment is echoed in the survey data (Figure 5.2 and Table C.8), with 49.6 % of participants in the base case (D1) expressing some level of discomfort with such a scan and only 39.5 % being (somewhat) comfortable with it. As expected, the more invasive variant (D2) was only viewed favorably by 19.3 % of respondents, while 26.9 % had a negative stance. If Daisy only sent pings from her own accounts (D3), these rates were 21.0 % and 33.6 %, respectively.

5.3.3.5 Eve – Path Traversal

Eve measures the prevalence of supposedly inaccessible files inadvertently exposed to the public, like path traversal vulnerabilities or exposed `.git` listings. This was the least controversial scenario among interviewees.

For criminal law, the interviewed legal experts focused on data espionage, as researchers were not manipulating any data in this scenario. The legality of scanning for confidential files depended on whether the researchers bypassed any access protection. Legal expert 3-L explained that Eve's action did not constitute the offense of data espionage, as *"the mere intention that something is secret is not enough to secure access; I need some objective barrier to access"*. Six other legal experts leaned in a similar direction, explaining that if the target file was hidden behind a random file name, like a UUID, this could be argued as an implementation of access control. However, if researchers were only looking for well-known files like *.git*, most agreed that this could not be considered an access control bypass. Regarding privacy laws, the experts repeated that researchers must have a valid reason for processing any user data: *"If it is personal data, then of course you have a processing of personal data. Consequently, it is somehow a data protection right and needs the usual justification"* (2-L). If no sensitive information is expected, there should be no privacy concerns.

The REC members in our study also saw the privacy aspect as the most critical one. However, they all agreed that such a scan would be acceptable if researchers did their utmost to minimize data processing. As participant 22-E put it in our interview, *"[W]hat I would try to do is try to develop a mechanism that minimizes the need for humans to look at data [...] that is sensitive."*

All but two operators we interviewed shared a similar perspective. Participant 13-O stated that the critical aspect was *"not knowing what happens to the information"*. On the other hand, 15-O said about the data: *"It is public! Whether I wanted it or not, it is just public, and now even the bad guys see that!"* All but two interviewed operators agreed that they would accept such scans.

In the survey, all Eve variants received similarly high approval (see Figure 5.2 and Table C.8): 67.2 % exhibited (some) degree of comfort with the scan in the base case (E1), while 21.8 % were (somewhat) uncomfortable with it. 41.2 % still answered

positively in the more invasive scenario (E2), while 27.7 % a provided an assessment that tended towards negative. The less invasive case (E3) did not lead to notably higher acceptance with initially skeptical operators; here 3.4 % where (somewhat) comfortable and 19.3 % still (somewhat) uncomfortable with Eve's scan.

Key Takeaways

Throughout all scenarios, legal experts often avoided making absolute statements, only leaning into one possible direction, underlining the complexity of the legal landscape. REC members occasionally referred to the legal side, yet focused more on users' privacy and considering alternative research designs, e. g., using dummy accounts. Operators mentioned a few red lines during the interviews, e. g., data leakage and modification, yet many were open towards 3S research, provided it is accompanied by a proper risk assessment and a responsible disclosure process.

5.3.4 Participants' Suggestions for Improvement

Throughout our interviews, legal experts frequently highlighted that 3S-based research currently occupied a gray zone, with benevolent researchers carrying the risks (7-L). Thus, most interviewees from law suggested legislative changes to add exceptions for security researchers (3-L; 7-L), akin to exceptions in drug research. 1-L also suggested a system for approving pre-registered projects. A counter-argument raised against both ideas is that the state cannot simply allow actions that could cause direct harm (6-L). In response to this, 5-La noted that in fact many potentially harmful activities, such as the operation of nuclear power plants, are already legally allowed. The more critical issue lies in defining who qualifies as a *researcher* to prevent the law from being abused by actors with malicious intent. However, legal experts also recognized that such legislative measures at the nation-state level would not extend to other jurisdictions, underscoring the need for a global solution.

The interviewed REC members proposed that security research institutions should establish their own ethical review boards or enhance IRBs with IT knowledge to raise concerns prior to conducting studies. Nonetheless, they advocated for ethical guidelines and procedures like *"ethics modeling"*, mentioned by 22-E with reference to literature [128]. These guidelines should not be strict, but rather outline a process that considers various ethical perspectives to address each stakeholder's interests (20-E, 21-E). In their publications, researchers should then discuss these considerations and justify how the chosen research designs minimized the potential for harm while still being able to answer the research questions. Lastly, interviewees admitted that there may be a lack of communication between RECs and the community, raising transparency concerns. They suggested that more information about the REC process and educational material such as practical ethics scenarios, like our vignettes, could address these concerns (20-E; 23-E).

Transparency is important not only for ethics boards but also for the operators in our study who desire clarity regarding scans performed on their systems. Measures such as identifying fixed IP areas from which scans originate (16-O; 20-O), including headers that indicate the scanning actor's identity (16-O), or fixed scanning time slots were

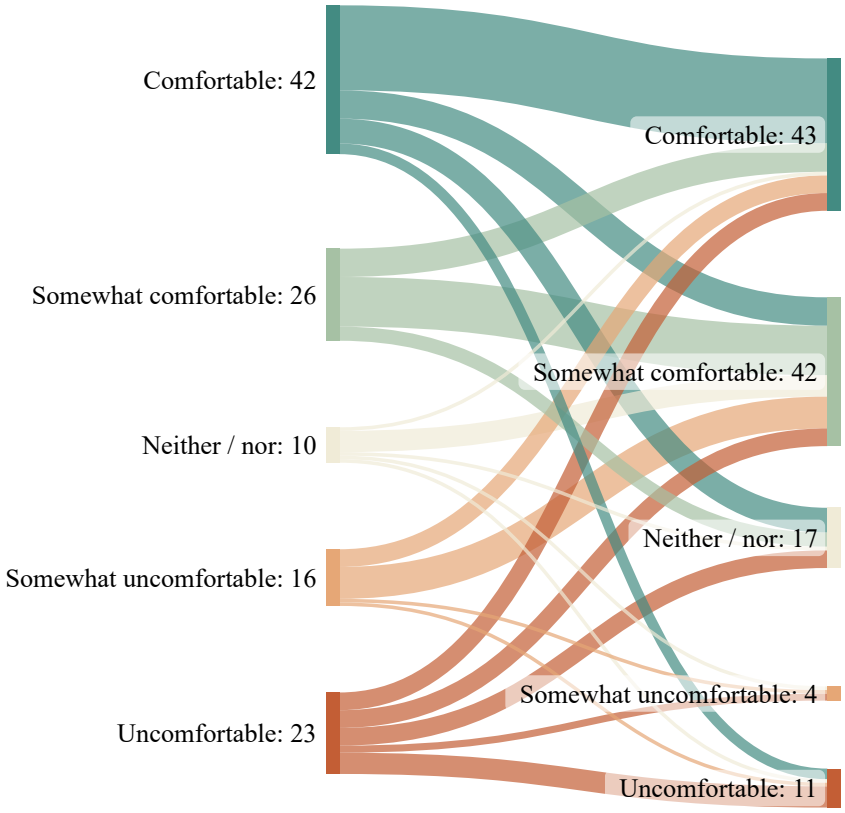


Figure 5.3: Changes in surveyed operators' stance towards 3S between Q6 (first general assessment; left) and Q8 (with pre-registration; right)

suggested as a means to differentiate between legitimate security scans and malicious activities (13-O; 16-O). Nevertheless, due to the challenges of justifying such measures to operators, almost all operators said they would prefer to be asked for consent or informed in advance about a scan. Then they would know who is behind a scan and could prepare their website, e. g., via *"declar[ing] clearly on the homepage that in the next two hours, in the next 24 hours, there may be errors"* (13-O). In general, the operators in our study wished to be actively involved in the process, and they consistently highlighted the importance of disclosing identified vulnerabilities responsibly.

5.3.5 Opinions on Pre-Registration

We also asked interviewees for their opinion on a pre-registration procedure overseen by a trusted third party (TTP), proposed in more detail in Chapter 5.5.2. Some had already suggested such a process themselves before we brought it up.

All academic legal experts we interviewed were positive about the idea, yet L-4 noted that pre-registration would not prevent civil lawsuits, e. g., in case of harm. Practitioners were less optimistic, pointing out that significant legislative change would be

required. Nevertheless, the general feedback was optimistic, with the German Federal Office for IT Security (BSI) viewed as a potential TTP. The REC members that we interviewed also welcomed the proposal but raised concerns about a national institution serving as the TTP due to the global nature of the scans. They would prefer a globally coordinated and community-driven initiative. The interviewed operators also disliked the idea of a governmental agency being in charge of a TTP, as they assumed bureaucracy would unnecessarily delay research. Instead, they favored an entity from the IT community (9-O; 18-O) or industry (9-O; 16-O). Still, the majority appreciated the idea of a TTP that would bring transparency to 3S research.

In the survey, we also asked participants about their general assessment of 3S; first before the scenarios and the proposal (Q6) and later in the context of our proposed preregistration with a TTP (Q8). Figure 5.3 compares the respective answers, detailed numbers are listed in Appendix C.4. Between the two questions, we observe a shift in opinion towards a more comfortable position. Those whose opinion changed from "comfortable" to a less positive position often explained this with unnecessarily hindering research (*"it doesn't make sense to restrict the security researchers"* [105-OS]), as malicious actors would perform scans anyway.

Asked to rate the competence of different entities to serve as a TTP in such a preregistration process (Q7) on a Likert scale from competent (score: 5) to incompetent (1), we find that the surveyed operators assigned higher competence to NGOs / white-hat hacker organizations (mean 4.15, std ±1.1, med 4.0) and academic institutions (mean 3.79, std ±1.13, med 4.0), while industry organizations (mean 2.67, std ±1.36, med 3.0), government agencies (mean 2.78, std ±1.38, med 3.0), and international organizations (mean 2.67, std ±1.36, med 3.0) were considered less competent.

5.4 Discussion

Our findings reveal numerous challenges when planning and conducting 3S studies. We now discuss the main insights from our interviewees and propose how to move forward.

5.4.1 Legal Roadblocks

Our project aimed to determine the legal boundaries of 3S research (RQ1). While we analyzed these boundaries through the lens of German law, conversations with international experts indicated some similarities in other countries. Among numerous aspects to consider from our interviews, we found three to be the most frequently raised.

No judicial decisions. Currently, there are hardly any judicial decisions on 3S or similar hacking-related research topics. Not only does this leave researchers with uncertainty about the legal situation, but it also makes it difficult for legal experts to assess how the courts would interpret the law and decide specific cases. On the contrary, our expert interviews left us with the impression that German legal practitioners deliberately try not to pursue such cases or to settle them out of court, apparently to consider the importance of white-hats and security research but ultimately stalling jurisprudential progress.

Need for legislative action. Given this lack of court decisions, even experts from

law struggled to provide clear answers to our legal questions. To provide researchers with a reliable regulatory environment, the interviewed experts suggested to introduce exemptions for researchers into German law to avoid prosecution and limit liability. However, this often raised the question of who such an exemption would apply to – only academic researchers, white hat hackers, and who else? Exemptions for security researchers are also of political importance, as underlined by this topic being included and continuously discussed in the past and current German coalition agreement [102]. The Netherlands [43] have already established similar exceptions.

International dimension. Although some countries already implemented legal exceptions for researchers, an international agreement is necessary, as the global nature of the Web and its vulnerabilities confronts 3S researchers with a multitude of jurisdictions. This further complicates the task for legal experts to provide reliable guidance regarding such scans. With modern web applications heavily relying on cloud hosting and content delivery networks, a server's location might be hard to determine and/or subject to change, making it hard to predict which jurisdiction applies.

5.4.2 Ethical Challenges

We sought to understand how RECs form their decisions and how 3S research can align with ethical standards (RQ2).

Ethical discourse. Our interviews revealed that REC decisions emerge from in-depth discussions with authors. Instead of having a preconceived opinion, they try to understand the authors' intentions and figure out the benefits of the research. They weigh them against potential for harm, which can be challenging to predict given the Web's complexity. The interviewees often stated that their aim was to help the authors rather than hinder them. For example, the REC at S&P 2022 gave a "Reject" recommendation only for two out of 67 papers flagged for ethics (from 1,006 in total) [29].

Missing guidelines. Our interviewees highlighted the absence of practical ethical guidelines. While the Menlo report [14] is a well-known standard, it is largely theoretical and lacks practical examples. Moreover, ethical understanding varies globally, suggesting that a single standard may be insufficient. Instead, we should learn from previous REC decisions, e. g., if published anonymized. REC members also repeatedly said that they disliked fixed rules for ethics due to edge cases and preferred to maintain their discourse approach to consider each case individually. They welcomed our 3s scenarios as a possible step towards creating ethical teaching material that fosters a fair discourse.

Post-mortem. REC decisions are made during peer review of a paper after the research has been conducted. This means that if an experiment went wrong, the harm had already been done, leaving the REC with no option to prevent it. Some interviewees emphasized they would prefer a mechanism to review research ex-ante. Others disagreed, arguing that this would entail too much work and that authors should be trusted. Examples such as the Tor Research Safety Board [242], medical research, or IRBs show that such a pre-approval of research projects is feasible if well designed [48]. We discuss this in more detail in Chapter 5.5.2.

5.4.3 Operator Perspectives

For our third question (RQ3), we wanted to get web operators' views on 3S, as they are the ones directly affected.

Legal action. Our interviews with operators revealed that most of them would not consider legal action against researchers for their scans. They see this only as the last option if direct communication fails to result in a solution or in an event of significant financial or reputational damage. Still, some operators might be obligated to file criminal complaints at the request of their company's legal team, regardless of the intention behind the scan.

Consent. Most interviewed operators were willing to support researchers, but to prepare for potential harm and know who to contact, many preferred to be asked for consent before the scan. However, 3S-based research typically needs to be conducted at scale for meaningful results. As shown by prior work [234] and the bounce rate of our survey invitations, reaching the operators of a large number of websites is a persisting challenge, let alone asking for their permission. Approval requests could also introduce observer effects and selection bias towards security-aware operators. As much as this approach would be favored, it is simply not practical.

Harm acceptance. 3S research has the potential to cause harm. Still, many interviewees stated that their log files already indicated such scans regularly occur in the wild. They do not expect additional harm caused by researchers. If researchers run non-invasive and well-pre-tested scans likely to get lost in the Internet's noise, the operators in our study might be willing to accept them. They further would find value in researchers responsibly disclosing vulnerabilities.

5.5 Recommendations

Our work confirms that server-side scanning is challenging and comes with the potential for harm. However, the majority of our interviewees agreed on the need to enable 3S-based research, as malicious actors also conduct such scans without prior notice but with ill intent. Based on our findings, we now introduce best practices for 3S research and discuss our proposal for a pre-registration board.

5.5.1 Best Practices for Server-Side Scans

Our interviews aimed to find out how 3S research can be conducted safely and without violations of legal or ethical norms. One core aspect was to first evaluate alternative research designs that also answer the research question but minimize the potential for harm, e. g., dummy accounts to test against instead of real users (see Chapter 5.3.3.3 for the suggestions and Chapter 6 for a practical application of this concept). If only a 3S-based design is viable, researchers should try to follow a few rules that partially extend prior recommendations [65] to ensure the benefits exceed the potential harm:

Laboratory pre-study. Before running the scans on the Web, there should be a laboratory pre-study. This way, researchers can catch potential issues early in the process and minimize the chance of accidentally crashing any servers.

Data minimization. The experiment should collect and store as little data as possible. Additionally, the process of validating data should be automated so that researchers minimize the risk of viewing personal information.

Data manipulation. Even though any request by definition triggers the server-side manipulation of some bits, researchers should limit data manipulation to the bare minimum and refrain from altering user-related data.

Resource minimization. The experiment should be designed to require as few resources on the server side as possible to not overload a server. If many requests are needed, they should be scheduled over a longer time period.

Monitoring. Researchers should always monitor the status and results of their scans. This means that not only should the scanner be monitored, but one process should also check that the scanned websites remain online after a scanning action. If a server crash is detected, the scan should be paused for closer investigation. For transparency, any incidents should be mentioned in the resulting publication.

Transparency. For the acceptance of 3S it is important to allow operators to learn about the purpose and scope of the scans occurring on their servers and the study goals. For this, we recommend creating a study website and linking to it in a custom header in all requests to the server. A reverse lookup of the scanner's IP address should lead to the same website. Some operators also would welcome a signature-based method to verify that scans really originated from the purported institution. We leave this to future work.

Fixed IP address. Not only should the scanner's IP address point to the study website, but it is equally important that the IP address(es) used for scanning remain the same throughout the entire study. This allows operators to add them to their allow list or opt out by blocking them.

Allow explicit opt-out. Operators should always have the option to explicitly opt out of the study. Thus, the study website should provide clear instructions on how to opt out and, ideally, a form to make this process as easy as possible.

5.5.2 Pre-registration Board

Based on our findings, we propose to establish a *pre-registration* process overseen by a *trusted third party* (TTP) for future 3S-based research. Although this approach cannot provide absolute legal certainty due to the global scope of 3S, it can offer proof of researchers' intentions, removing some subjective elements in laws such as Germany's §303b StGB on computer sabotage.

The proposed TTP would review submitted research proposals that should outline the research question, expected outcomes, and the chosen study design, as well as a discussion on why alternative methods are not feasible. This will allow the TTP's ethics reviewers to weigh the potential for harm against the benefits, engaging in dialogue with the researchers. Ideally, the TTP should be supported by legal advisors, given that ethics reviewers cannot be expected to know the applicable laws of all jurisdictions.

To avoid excessive simultaneous testing of systems, the approval of research proposals should include the assignment of multiple time slots, during which the scans can be performed. The TTP should also maintain a publicly available list of all ongoing

scans, the associated IP addresses, and points of contact, enabling operators to get information about current approved scanning activities.

As our survey suggests, the TTP should be established as an NGO, possibly with support from academics from all over the globe to represent diverse ethical views. While we are aware that establishing such an entity is an enormous task, we believe it would greatly benefit future research.

5.6 Summary

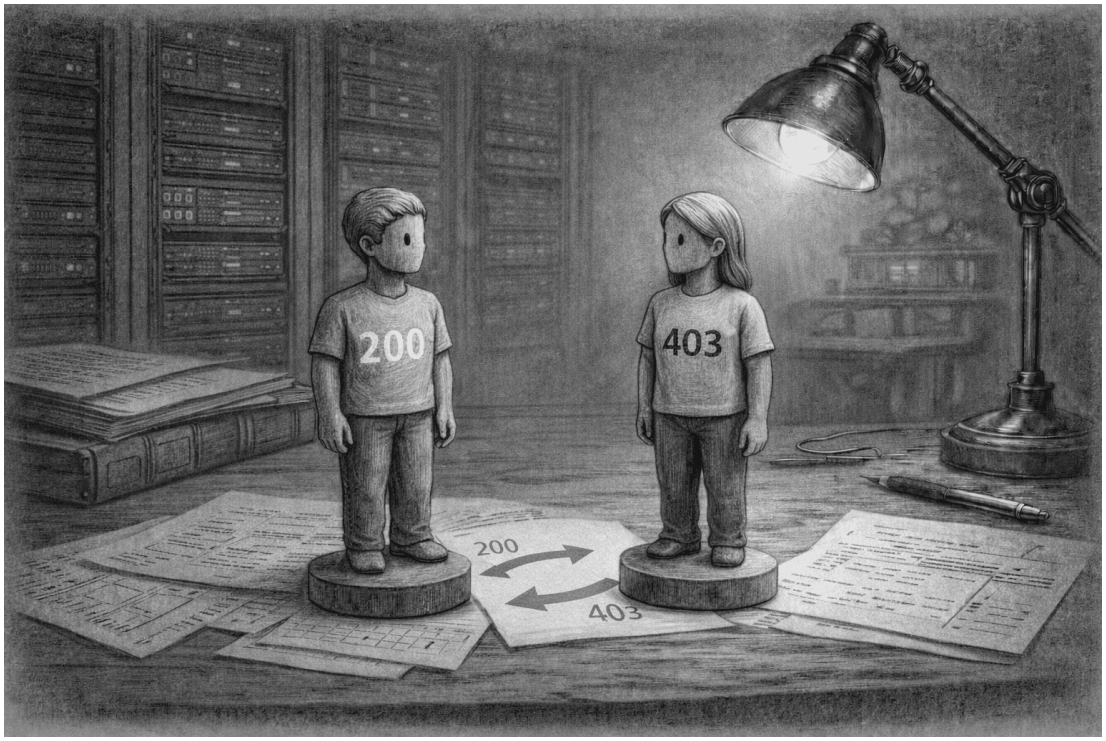
We explored the normative boundaries of security and privacy research based on server-side scanning through interviews and a survey. While ethical and legal risks for researchers have led to reluctance in conducting server-side scanning research, our findings reveal that the slight majority of operators in our study might be open to server-side scanning, yet there is a need for global legislation and ethical guidance early in the research process to provide researchers with a clear and reliable framework. Our hope is that the best practices and pre-registration approach proposed as the outcome of our study will provide guidance toward this goal and stimulate productive discourse within the security and privacy research community.

As this topic is also a political one, we also hope to raise awareness beyond academia and to provide more normative clarity to non-academic researchers and ethical hackers around the globe. We encourage other researchers to build upon our work and conduct similar studies in their own countries. This would allow our community to gain a more comprehensive understanding of the normative boundaries of server-side scanning and security research on an increasingly larger scale. The envisioned result could be a global set of research best practices and, ultimately, more research-friendly adjustments of policy decisions worldwide.

Taking the insights from this chapter, we present in the next chapter an example of their practical application by analyzing and scanning for a widely known server-side issue, access control issues.

6

Measuring Server-Side Access Control Issues: An Ethical Case Study



Contribution of the Author

This chapter contains parts of the papers *"403 Forbidden? Ethically Evaluating Broken Access Control in the Wild"* [S1] which was a collaboration with Saiid El Hajj Chehade, supervised by Ben Stock. The group designed the experiments together. Florian Hantke led the ethical considerations of the work, while Saiid El Hajj Chehade implemented them and performed most of the analysis, with minor support by Florian Hantke. Florian Hantke also conducted the responsible disclosure for the study. All authors commented on the text and contributed to individual sections.

6.1 Ethically Evaluating Broken Access Control in the Wild

In the previous chapter, we learned about legal and ethical challenges associated with large-scale scans of server-side issues and how to deal with these challenges. In this chapter, we present a practical example of how to conduct ethically ambitious research on the Web in a responsible way, following the insights from earlier.

By now, it is clear that the Web is widely used and indispensable for everyday tasks. However, web applications are becoming increasingly complex, and with the rise of AI-generated code, projects may become harder to maintain. This complexity can easily lead to Access Control (AC) issues, a server-side vulnerability listed as the number one issue in the most recent OWASP Top Ten of application vulnerabilities [177].

In this chapter, we evaluate a method for measuring the prevalence of *AC* server-side issues at scale. In particular, we draw on discussions from the Charlie IDOR case study in Chapter 5, where interviewees suggested using two controlled accounts to execute attacks between the two. To follow this advice, we implement a framework that uses a leader/follower approach to browse applications with two researcher-controlled accounts simultaneously, ensuring mirrored user interactions, and filter network requests to identify HTTP endpoints with potential broken AC. For each such endpoint, we swap resource identifiers between the two users' requests. Based on heuristics, we then manually assess flagged cases. To summarize, we make the following contributions:

1. We conduct a thorough analysis and discussion of research ethics when scanning for server-side access control vulnerabilities, highlighting how to minimize risks and maximize benefits for the entire ecosystem.

2. We design and implement the Variable Swapping Framework for conducting ethically sound access control vulnerability detection in the wild.

3. We report on the results of our experiments and highlight that out of the 100 sites we tested, six are susceptible to at least one broken access control bypass, threatening users' privacy.

This chapter discusses only the parts of the study most relevant to this dissertation. For a more detailed description, we refer the reader to the original paper [S1].

6.2 Ethical Considerations

Building on the earlier insights, we know that for an ethically challenging project, we first conduct a stakeholder analysis (see the Menlo Report [14] and Kohno, Acar, and Loh [128]) to systematically evaluate the ethical implications. It helps to develop guidelines to adhere to throughout the project.

6.2.1 Stakeholder Analysis

When investigating *AC* issues, three main stakeholders are involved: the end users of the website under analysis, the website operators, and the research team itself.

End Users End users are those who interact with the web service. Although they are only passively involved in the experiment, they still face potential risks. A poorly designed *AC* experiment could expose sensitive user data to researchers or, in the worst case, modify it. Activities that interact with the end user, such as unsolicited messages, can also negatively impact the end user by causing stress, annoyance, or disruption. If the scans are too aggressive, they may also affect the performance and availability of the web service, leading to a negative end-user experience. We discuss measures to minimize, and ideally eliminate, these risks in the next section.

Despite these risks, we believe the benefit to end users is significant. Each vulnerability that is discovered by the research team and fixed by the website operators improves the web service's security and reduces the risks of end-user data being leaked. By minimizing the risks outlined above, we argue that the overall benefits to end users outweigh the potential risks to individuals. Data leaks, if exploited by malicious attackers, could affect all users of a website, whereas measurement experiments can be carefully designed to minimize interactions with users' data.

Website Operators Unlike end users, website operators are actively involved in the experiment. For that reason, it must be discussed whether and how to seek their consent, and if not, to justify why [14]. As explained in the previous chapter, seeking consent from operators for a large-scale web measurement is not feasible. However, in *AC* experiments, operators may be affected by additional traffic, leading to higher costs. In the worst case, their website's stability suffers from the additional traffic. Creating fake accounts to test *AC* issues can add additional administrative overhead. Furthermore, exposed user data due to a poorly designed experiment may harm operators by causing stress and additional work to triage the incident and inform affected users. Public disclosure or naming the web service in publications could also pose a risk of reputational damage. At the same time, the benefits for website operators are substantial. If a vulnerability is identified and disclosed responsibly, it helps operators understand and address *AC* weaknesses on their site. Ultimately, this helps to improve their service's security, reducing the risk of costly data breaches.

Researchers and Team Members Finally, the research team conducting the experiment is also a stakeholder. While all team members have consented to conduct this

research, it is still necessary to consider and minimize potential risks they may face. As with any web-based research, there is a possibility of encountering content that could be uncomfortable. Additionally, researchers may face legal threats from parties who do not approve the unsolicited web scans.

The benefit for the research team is eventually a contribution to the academic community through the publication of their findings, which will raise awareness and understanding of *AC* issues. Furthermore, like other end users, the researchers benefit from any positive security improvements resulting from this work.

6.2.2 Ethical Guidelines

To address the list of potential risks identified in the previous analysis, we discuss each risk in the following and propose measures to incorporate into our framework for detecting server-side access control vulnerabilities.

Data Leakage and Manipulation *End users* and *website operators* could both be negatively affected by leaked or manipulated data. In Chapter 5, we interviewed (among others) REC members from major security conferences. As a result of the discussion, the REC members suggested that a good way to avoid exposing end-user data is to experiment only with accounts owned by the research team. This way, any potentially leaked or updated data would only affect the respective other researcher-owned accounts. The framework must thus ensure that it interacts solely with researcher-controlled accounts. While it is possible that scans may view publicly available user information, the framework ensures that no sensitive data is accidentally disclosed. In addition to the technical measures, all researchers who are involved in the vulnerability analysis agreed on a self-commitment declaration not to use the gained knowledge to act outside the experiment context.

Fake Accounts As mentioned above, the framework ensures that we only test with researcher-associated accounts. Due to the administrative overhead that fake accounts could impose on *website operators*, a balance must be maintained in the number of accounts created and required. We decided to create two accounts per site that have recognizable test account names.

Informed Consent As mentioned earlier, it is essential to discuss whether and how to seek informed consent of the *website operators*. Ideally, researchers would request consent from every web operator whose website is tested. However, research has shown that even vulnerability notifications receive a very low response rate [234], and we expect even lower rates when asking for consent, limiting the study to a very small and potentially biased scale. According to the Menlo Report [14], research that would otherwise be infeasible can be performed without an informed consent process but must be well justified. This is the case for our research, which aims to design a framework for a scalable study that is indeed infeasible if operator consent is required. Hence, we have chosen a study design in which we do not acquire operator consent. To ensure this

design aligns with current research ethics, we consider ethics in every research stage to minimize risk for all involved parties and have obtained approval from our ERB.

To still give the operators the right to withdraw from our research [14], the framework should include information on how to withdraw (in our case, in an HTTP header) and use one fixed IP address that can be blocked.

Website Resources and Stability An immense usage of resources or threat to the website's stability due to web scans would have a negative impact on *end users* and *website operators*. To minimize this risk, the framework should only send a normal amount of valid web requests, similar to those made by the typical user. Additionally, we decided that every request that is automatically sent with our framework needs to be successfully performed manually by one of our researchers in the form of visiting the site. For unexpected responses, e.g., a 500 Internal Server Error, during a later automated stage, there must be safeguards in place to react accordingly. In our framework, we see such responses in the database so that we can investigate the difference further and verify the page's availability.

To avoid additional costs or impacts on the website's stability with too many requests, the experiment framework should enforce request limits per site, ensuring that the generated traffic remains negligible compared to the already existing noise on the Web. We decided to limit the requests to 16 per HTTP endpoint.

Alarms Same as with the resources, potentially triggered alarms could put stress on *web operators*. At the same time, any alarm that leads to the earlier detection of an existing vulnerability – even before we disclose it – is a good alarm, as the vulnerability should not exist in the first place. Such vulnerabilities, if exploited by adversaries, could lead to mandatory reporting under privacy regulations (e.g., GDPR), causing even more work or even financial penalties for the operators. By ensuring that the scanning framework only uses researcher-controlled accounts and avoids leaking any sensitive data, the risk and the additional work are minimized for the operators. Furthermore, a request limit implemented in the framework lowers the number of potential alarms.

Responsible Disclosure The key benefit that *all stakeholders* get from this work is the identification and remediation of detected web security issues if disclosed properly. Therefore, a disclosure strategy must be considered before the experiments begin and be balanced with the timeline of the experiment. Issues disclosed too early could influence measurement results (e.g., operators fix the same issue for multiple of their sites in our dataset), while issues disclosed later could raise the risk of exploitation by malicious actors. Therefore, we decided to prepare our disclosure communication during the execution of the experiments and send it for every verified vulnerability immediately after all scans are finished. The disclosure emails must contain enough resources and recommendations to minimize the additional work that web operators do.

Uncomfortable Content To protect *research team members* from viewing content they feel uncomfortable with (respect for persons [14]), we ensure that team members

have the option to skip a website at any time. In such a case, the lead researcher would review and evaluate the skipped site.

Legal Issues Legal risks always exist for a *research team* conducting this type of research. We have carefully considered our experiment pipeline and are confident we do not violate any legal regulations, as our testing is only limited to two researcher-controlled accounts, thereby avoiding the classical brute forcing of IDs and parameters. One exception is creating test accounts, which may still violate some websites' terms of service. We accept this risk as appropriate, as we will not pursue further action if operators block or delete our accounts. To further minimize legal risk, we run all experiments from our institution's IP address rather than from private IPs.

Summary Broken AC vulnerabilities can not only cause financial loss to the operators but also jeopardize the privacy of millions of users, as evidenced by the previously mentioned examples. As noted in Chapter 5, ethics committee members, as well as the website operators, were generally confident about such kind of research when conducted responsibly. Additionally, we carefully studied and discussed the recommendations from relevant ethics research publications [14, 128], and requested feedback from our ERB. They concluded that "[t]here are no ethical concerns against the implementation of the proposal"; however, they suggested that the involved researchers sign a self-commitment declaration promising to not "use the gained knowledge to act outside the experiment context". Additionally, they mentioned that all involved researchers should be made aware of potential legal actions that vendors could take and the measures we have implemented to mitigate these to enable an informed decision about whether they wish to participate in the research. We have followed both of these recommendations.

In summary, given the significant risks of data leaks from improper access controls, it is paramount to identify such flaws and inform operators to have a positive impact on the security of the entire ecosystem. Thus, we believe the potential risks to stakeholders outlined in this section are significantly outweighed by the benefits of identifying such flaws, assuming the research framework is designed to minimize risks.

6.3 Threat Model

In this work, we study attacks that exploit flawed server-side AC policies to gain unintended privilege over some or all user-accessible permissions or resources in a web app. Specifically, we consider scenarios where the attacker already has their own account and abuses AC policies to gain access to or manipulate a victim user's resources by altering request parameters [53]. The general pattern for such attacks is replacing the identifier of a resource belonging to the attacker with the identifier of a victim's resource. If the application fails to properly verify authorization, the server may return or modify the victim's data. In many cases, such identifiers can be changed easily, e.g., by editing the URL or sending requests from the browser dev-tools [52, 171].

Our threat model therefore focuses on adversaries who obtain valid resource identifiers – either through guessing, observing publicly available information, or interacting with the victim – and exploit improper authorization checks. Unlike account takeover

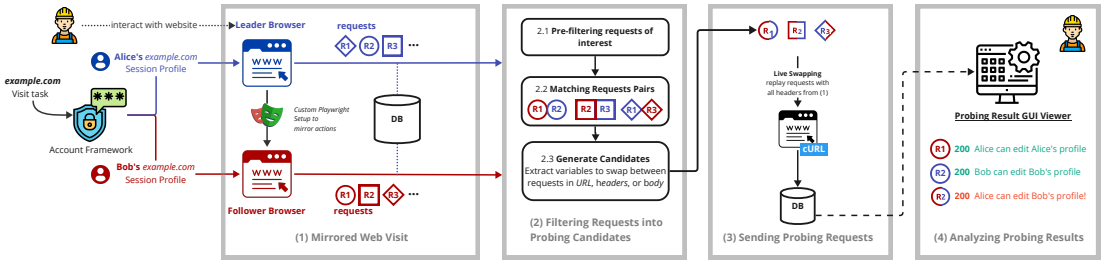


Figure 6.1: The Variable Swapping Framework Simplified Workflow

attacks, the attacker does not require the victim's authentication credentials and instead relies solely on manipulating identifiers in requests. Consequently, the primary defense against such attacks is the correct implementation of server-side *AC* policies rather than relying on the secrecy or complexity of resource identifiers.

6.4 Methods

Without access to backend source code, evaluating whether web endpoints implement sound *AC* in a black-box manner is not straightforward. First, studying *AC* issues is only relevant to websites with logged-in user accounts that can own private resources (e. g.profile pages, files, etc.). Second, challenging the server's *AC* policy requires attempting to request other users' resources within a user's session. Throughout this section, we refer to requests which aim to find improper *AC* checks as *probing requests*. While successful requests indicate a faulty *AC* policy, they might cause an ethical violation by revealing other users' private data. As stated earlier, to avoid compromising external users in our study, we must only target users we control, meaning we need at least two users per website. As managing all the above by hand is not scalable, we design the Variable Swapping Framework (VSF) framework(see Figure 6.1)), which automates most parts of the experiments, leaving only peripheral human involvement. To efficiently manage user accounts and instantiate authenticated website sessions – in a semi-automated fashion – we adapt the *Account Framework* by Rautenstrauch et al., which distributes fresh logged-in sessions to our automated browser agents. To generate probing requests ethically, we collect pairs of requests from two fake users logged in to two browsers – a *leader* browser controlled by the researcher who performs usual user interactions and a *follower* browser that mirrors actions from the *leader*. We then introduce an automated pipeline to match request pairs and extract *AC* parameters (user credentials, resource identifiers, etc.). The VSF then constructs probing requests by only swapping the resource identifiers between requests while keeping the same user credentials. We deploy automated browser agents to send these requests and record the responses in a central database. Finally, we review and analyze flagged requests to manually confirm vulnerabilities.

6.4.1 Semi-Automatic User Accounts Management

We integrate the *Account Framework* [189] to store and manage the user accounts for the various websites. As a first step, we perform manual registration through the provided assisted setup to create two user accounts per website, keeping track of e-mails, usernames, and passwords. All users share one of three e-mails we control. Rautenstrauch et al. used the created accounts in a fully automated fashion by crawling the sites through following links. They did not, however, interact with the page further (e.g., triggering forms or buttons). In contrast, we must perform certain unskippable actions (e.g., filling out the profile information on the first login) in *both* accounts. For this, we develop a protocol to ensure that our two accounts are in the same state before we start interaction with the *leader* browser.

The *Account Framework* then automates the rest of the process to serve fresh sessions for the VSF (Figure 6.1), by automating any required logins for expired sessions. We refer the reader to their paper for a detailed explanation of the inner workings of their setup.

6.4.2 Manual Mirrored Website Visit

For each endpoint to evaluate *AC* behavior, we need examples of accepted resource identifier values for both users (e. g.folder ID), so that we can later swap between users (Chapter 6.4.4) to ensure only researcher-controlled is accessed. Effectively, we must collect an example request for the site endpoint from both user sessions. Naturally, we cannot claim to exhaustively evaluate all endpoints for a given site; we rather restrict ourselves to covering the *same* set of endpoints between two sessions to have the necessary example requests in both cases.

To efficiently explore available endpoints and collect two examples from two distinct user sessions, we design a mirrored browser setup, as shown in Figure 6.1 (1), maneuverable by a single researcher. The researcher controls one browser, labeled *leader*, and navigates the website with the first user account (Alice). A second browser, labeled *follower*, automatically mirrors each interaction (clicking, filling inputs, navigating URLs) produced by the *leader*, but with the second user account logged in (Bob). Intuitively, synchronizing the browsers' interactions guarantees a higher likelihood of generating two examples of each discovered endpoint from the two user accounts. We collect request/response pairs from each browser in a database for later stages of the experiment. In case the follower browser falls out of sync, the researcher can directly interact with it to bring it back to the same state as the leader browser.

6.4.3 Filtering Requests into Probing Candidates

We then filter the requests. This step's goal is to minimize the number of irrelevant requests and useless swapping experiments, reducing the analysis workload and the network impact on the tested server. We disqualify requests and probing candidates in three stages, visible in Figure 6.1 (2).

We first pre-filter requests to only keep candidates relevant to access control. This means we drop requests that match the popular Ad-blocker lists, exclude static re-

sources (e.g., CSS), and remove all requests that do not indicate any form of authentication, such as cookies or authentication headers. Also, we remove request/response pairs that are identical across both sessions.

Because the two mirrored sessions can produce misaligned request sequences, in the second step, we match requests using a URL path similarity heuristic.

In the third step, we generate from each matched pair probing candidates by applying a templating process which takes one user's requests and identifies the variable parameters likely to be resource identifiers (see details in [S1]). We then have a template request that can swap the parameters from one request to another to generate probing requests.

Finally, we can generate multiple probing requests if the reference request has many resource identifiers. From the set V of available identifiers, the candidate generation module selects a subset of variables V_S to swap based on heuristics. Then, it can generate all $2^{|V_S|}$ combinations. To avoid overwhelming the server and keep the time spent per site short, we put a hard limit of 16 candidates per endpoint.

6.4.4 Sending Probing Requests

Next comes the swapping module. Its key requirement is to ensure that the difference in server-side responses to probing requests compared to visitation requests is only due to the swapped parameters. The main factor that can cause the server to respond differently, besides the swapped parameters, is expired security parameters: Authentication credentials (e. g., session cookies, JWT tokens, etc.) can expire if the swapping experiment is sufficiently delayed, causing the server to respond with an unusable authentication error response. We guarantee this *parameter freshness*, by running a live swapping experiment in parallel to the visitation. On a rolling basis, the swapping worker fetches probing requests generated by the prior module. This approach guarantees that short-lived session parameters – like CSRF tokens – remain valid and do not influence the response to probe requests.

6.4.5 Manual Candidate Validation

In the last step, we use the GUI viewer to understand whether (1) successful probing requests (with status code 200) present actionable vulnerabilities and (2) failed probing requests failed due to AC policy violation or orthogonal reasons.

Ideally, we expect the response to the swapped AC rule to return an HTTP error code. As such, successful responses to probing requests are markers of likely vulnerabilities. To validate whether such candidates violate the expected AC policy, we perform a series of manual checks: (1) we contextually determine whether the swapped parameter is a private resource identifier. If yes, (2) we compare response bodies: if swapped requests body is closer to the one of the other user (which shares the same resource identifier) than for the initial user (which shares the same credential), we consider it a **violation of the *AC* policy**, as user the initial user was able to access the other users's resource. In case no response body is available (for `POST` requests for example), we consider successful requests over private resource identifiers a **violation of**

the *AC* policy. While this can be automated [191], we manually confirm that the improper authorization is an exploitable vulnerability, so that we do not notify operators incorrectly.

6.5 Results

We use the framework described before to scan web application endpoints for *AC* vulnerabilities and observe how servers respond. As the focus of this chapter is on the ethical methodology of the study rather than its empirical findings, we briefly summarize the key results in the next sections and refer readers to the published paper [S1] for more detailed findings.

6.5.1 Website Selection

Most stages of the framework scale to any number of websites. However, we had to limit the number of visited sites due to the manual effort required to register accounts and visit websites. To still cover sites relevant to users, we sample targets from the CrUX dataset [200]. The Account Framework [189] detected 531 websites across the 0-5K range as having login and registration pages, of which we could create account pairs for 110 websites between August 2024 and October 2024.

6.5.2 Testing Web Endpoints in the Wild

This section showcases the results obtained from visiting live websites with the VSF. In Table 6.1, we provide an overview of the relevant counts of websites and requests throughout the various stages of website selection and the VSF pipeline that we describe below.

Visiting Websites and Interactions We intentionally do not attempt to automate website visitation and interaction. This manual step ensures unintended unethical interactions cannot occur (e. g., interacting with a real user). It would be challenging to give ethical guarantees without tightly controlling the interaction footprint of our users by hand or sacrificing interactions altogether, like prior attempts [191].

Visitation Guidelines For each website, we enforce a clear visitation protocol. In essence, we instruct the researchers visiting the website to perform interactions that would *(A) expose AC-related endpoints* (e.g., navigating/editing the user profile, manipulating user-owned objects, liking/bookmarking public resources, etc.) without *(B) compromising the ethical bounds* (e.g., not reporting other users). All subsequent stages are automated and ensure that parameters making it to the probing requests are only pertinent to one of our two users, not harming any third-party user.

Failed Visits Out of the 110 websites we visited, we marked 100 as successful. Out of the 10 failed visits, three websites, marked as logged in by the Account Framework

Table 6.1: General experiment statistics from visitation to probing requests

Website Statistics	
Websites with registration and login forms	531
Websites with two accounts	110
Successfully visited websites	100
Websites with probing candidates	90
Visitation Request Statistics	
Total requests collected	536,973
Unique endpoints collected	60,153 (11.2%)
First-party requests	24,352 (40.4%)
Third-party requests	35,801 (59.6%)
Unique requests dropped by ad blocker	27,609 (45.9%)
Pre-filtered and matched endpoints	7,311 (12.2%)
Probing Request Statistics	
Request to swap	1,566 (2.6%)
Total probing candidates	6,830
Successful probing candidates	6,096
Unique probing URL templates	584
Endpoints improperly handling disallowed AC	30
Vulnerabilities	19

(Chapter 6.4.1), would not provide logged-in sessions during visitation. Further investigation shows our accounts were either blocked and required an account recovery process, or bot detection invalidated our sessions. Three other websites failed because we could not have any meaningful interaction. We further outline interaction limitations in the paper [S1]. The remaining websites failed because the web pages were too dissimilar for users or were very unstable. The three main reasons for session divergence between users are (1) AB-testing where a different user interface is selectively shown to one portion of the users e. g.how the share page responds on `notion.com`, (2) randomized content in the landing page e. g.images shown on `pinterest.com`, and (3) difference in user account states e. g.one user getting a random achievement on `boardgamearena.com`. While the VSF falls back to clicking at the same position on the page if the interaction target is missing in the *follower* browser, differences in UI layouts will quickly lead both browser sessions to become out of sync. Finally, we note that some websites, primarily adult content sites, use aggressive ad techniques that open many new tabs on every click, rendering a coherent visit infeasible.

6.5.3 Authorization Vulnerabilities

With the VSF, we collected more than 500,000 requests for which our heuristics (see the details in the paper [S1]) identified 1,566 to be requests for which identifiers can be swapped (see Table 6.1. After deduplication of request URLs, we ended up with

584 candidate URL templates, for example, `/users/alice1111/1234.json` and `/users/bob2222/5678.json` which are reduced to the candidate URL template `/users/<userID>/<file ID>.json`.

After an in-depth manual analysis of the flagged endpoints (described in Chapter 6.4.5), we found 30 out of the 584 (5%) endpoints to improperly respond to requests with bad AC pairs. They span across 15 sites out of the 100 tested. While improper responses are not necessarily immediate vulnerabilities, they can be a strong indicator that AC checks were not properly implemented, which developers could unknowingly transfer to other (more sensitive) parts of the application. Our exploit analysis concludes that 19 out of 30 improper responses are exploitable vulnerabilities with tangible harms: leaking private user information, partial credential hijacking, and user resource manipulations. The remainder of broken AC responses either (a) return successful responses, but the responses themselves are not valuable (e.g., returning post recommendations) or (b) have resource identifiers that are ephemeral or impossible to leak (e.g., image hash for Amazon storage services). We enumerate here the main (non-exclusive) causes certain AC responses qualify as vulnerabilities: using identifiers with low entropy (e. g.incrementable, timestamp and incrementable, etc.) – such identifiers can be easily manipulated by brute forcing identifieres (5 vulnerabilities); supplying identifiers that can be inferred from the user credentials in a separate field, e. g., email in the request body (1 vulnerability); using the same resource identifier for private endpoints and in endpoints accessible to other users or the public (e. g., URL leaks [173]) (13 vulnerabilities); and high entropy resource identifiers without any AC checks (6 vulnerabilities).

In the following, we describe three serious vulnerabilities we found as examples of instances of flaws. We have disclosed the issues to the vendors but have not received confirmation of fixes nor explicit consent to name to some of them, hence we redacted some details.

Vulnerability 1: RedactedNews.com The first service we examine is a news website that allows users to make donations on one of its pages. This page initiates a request to */api/payment-provider/user*, containing the user's mail in the body JSON. In response, the server returns an access token for a payment provider, which can be used to receive the user's payment information and purchase history. However, anyone can input any user's email address to obtain the other user's access token, thereby gaining unauthorized access to that user's payment information.

This unauthorized access could be exploited by a *targeted attacker* who first needs to know a valid user email address. However, since it is also possible to brute-force common email addresses or use leaked email lists, a weaker attacker model could also apply here.

To address this issue, we recommended that the vendor validate the authorization via the session user's email instead of relying on the JSON body input. Any input that could be manipulated by an attacker must be handled with caution.

Vulnerability 2: RedactedSocial.com The next service is an adult dating website. Every profile can be previewed as a profile summary with the information received via

the API endpoint */coreapi/profile_summary?handle=<user handle>&id=<profile id>*. This can be done using either a user handle or their user ID. While the handle is visible on the website, the IDs are unique numeric identifiers used only by the API. The ID consists of two parts, totaling 12 digits, and appears only partially random. Retrieving someone's profile summary remains possible even if the corresponding user profile was set to invisible for other users.

An opportunistic attacker could easily enumerate the IDs to gather user information about several users who intended to be hidden. An attacker could also target a user by directly using a specific user handle.

This issue stems from the API's failure to verify whether a profile is set to be invisible. We, therefore, recommended that the vendor implement this check within the API to resolve the issue.

Vulnerability 3: RedactedEditor.com The third vulnerability we describe was found in a PDF signature service. A user can upload a PDF and distribute the document to other signers via mail. For a new signature task, a task token is created. Visiting `/signature/request/<token>`, the user can see and configure the signature request. Using the same token with the API `/v1/signature/user/<token>` further reveals information about the signers and files. Although the signature task is bound to accounts via email addresses, every user possessing the token can open the request, download the linked files, and view all signers' info.

The token itself is 70 random characters long, so presumably very secure. Since the signature requests are sent via e-mail (which could potentially leak the token as we showed in related work [S2]), this is a prime example of a vulnerability that falls in an attacker model putting efforts in opportunistic reconnaissance.

To resolve this vulnerability, we would recommend verifying the request's authorization by checking both the session user's email and the signers' emails. Relying only on a randomly generated token in the URL is not enough.

These three vulnerabilities were discovered due to unique design features of the VSF: First, vulnerabilities (1) and (3) require sending `POST` requests, which prior work avoids due to ethical concerns; the VSF can send `POST` requests because we are sure to manipulate our own users' data only. Second, vulnerability (3) requires a complex setup of specific resources and user states (e. g.creating a signature request), which is only possible ethically with controlled manual visitation in the VSF.

Responsible Disclosure We reported all 19 vulnerabilities to the relevant parties via the most appropriate security channel available, including bug bounty programs, privacy email addresses, and generic email addresses such as `contact@` or `help@`. We reported the issues in early November 2024. By April 2025, we had received responses from three operators, each acknowledging our findings, with one providing a bounty payout. Upon revisiting the issues in April, we found that two of the operators had successfully resolved them, while work remains to be done by the others.

6.6 Summary

This chapter demonstrates that, on the one hand, access control vulnerabilities are indeed a relevant issue in the wild. In the study, we find that 15% of the sites incorrectly handled bad AC requests, with 6% actually being vulnerable a number that is not negligible. This confirms the high rank of the access-control vulnerabilities in OWASP's Top 10 [177]. Most issues occur because of redefining the identity through variables other than the credentials, e. g. having a separate `username` field, or because of relaxing the need to maintain AC on resources referenced by long and complex identifiers. Even when AC responds correctly, the variance in how it reports forbidden access varies widely between sites, making it harder to automate the analysis, e. g., returning HTTP 200 responses but containing an error JSON object in the body.

At the same time, this chapter illustrates that ethically ambitious server-side web security measurement research in a real-world environment is ethically possible if ethical considerations are integrated into every stage of the research process. As mentioned earlier, frameworks such as stakeholder analysis help researchers to deeply assess the potential risks and benefits their work may present to individuals and encourage them to really think about their place in the context of society, fostering more responsible, refined, and thus rigorous and trustworthy research. While these principles may appear abstract, the study presented in this chapter is a good example of how they can be applied in practice.

However, these considerations and principles are all for nothing if researchers are not (made) aware of them. That said, in the next chapter, we take a closer look at top-tier conferences and how they foster ethical research practices.

7

Ethical Practices at Security and Privacy Conferences

Contribution of the Author

This chapter is based on "Reflection, Education, Consistency: Towards Best Ethics Practices At Security And Privacy Conferences" [P4]. Florian Hantke, Rafael Mrowczynski, Til Dralle, and Ben Stock conducted the study. Florian Hantke developed the initial idea of this research and refined the methodology together with Rafael Mrowczynski. Both conducted and analyzed the interviews. Ben Stock supervised the study. Til Dralle experimented with trend analysis. All authors contributed comments and text to individual sections.

7.1 Motivation

As we have seen thus far, web security and privacy measurement research can raise serious challenges that require ethical considerations. However, if ethical considerations are not properly motivated or enforced by the conferences to which researchers submit, such considerations may not always receive sufficient attention.

Past cases illustrate the potential consequences. Controversial cases such as the Encore paper [36] or the University of Minnesota study [95] have raised discussion within and outside of the S&P community. These cases highlight ethical concerns with security and privacy research that go beyond traditional human subject studies, extending into areas such as vulnerability research, implying coordinated disclosure, and others with potential harms caused by research activities or outcomes.

Conferences have, in fact, responded to these incidents and introduced ethics standards, Research Ethics Committees (RECs), and increasingly often demand sections describing ethics considerations. Despite these advancements, the top-tier conferences – USENIX Security, IEEE S&P, ACM CSS, and NDSS – are still refining balanced and consistent ethics standards. Currently, their approaches remain unaligned, and differences in requirements can create inconsistent expectations across venues. For example, USENIX Security introduced mandatory ethics sections in 2025 [16], which caught many authors off guard and led to an over 20% administrative rejection rate in its first cycle, primarily due to missing ethics sections [17]. Such cases illustrate the need for more consistent and clearly communicated ethics standards.

One key challenge in achieving such consistency is the absence of well-defined and explicitly communicated goals and shared understandings of what ethical standards should achieve. Defining a set of goals is essential to evaluate existing procedures, identify areas for improvement, and move the ethical debate forward. Such alignment also supports more consistent and transparent decision-making across conferences, reducing the uncertainty for researchers of unknowingly violating policies and being rejected.

We fill this gap in research by defining goals and evaluating implemented ethics interventions. We first conduct a qualitative content analysis of both call for papers and review forms, and combine the findings with 20 semi-structured interviews with senior and junior members of our community. This collection of data allows us to explore how ethics practices evolved over time, what motivated their implementation, and how they help to achieve the identified goals.

Specifically, we answer the following research questions:

- RQ1: *What do junior and senior members of the S&P community perceive as the primary goals of ethics-fostering practices at major (top-tier) S&P conferences?*
- RQ2: *What ethics-fostering interventions (a) are currently in place, and (b) which additional interventions are desired by members of the S&P community to meet identified goals?*
- RQ3: *How can conferences improve their ethics-fostering interventions in accordance with the goals and needs as perceived by community members?*

Resulting from this, we present four sets of goals and find that while interventions address external impact and perception, community awareness still remains an area for improvement. While we view mandatory ethics sections (with minor adjustments) as a step in the right direction, we advocate for shared, community-wide ethics policies and educational material coordinated through an ethics steering body.

In summary, we make the following key contributions:

1. We present a qualitative analysis of ethical-reviewing practices and their development over time based on a content analysis of 99 CfPs and 20 semi-structured interviews with members of the S&P community.
2. We provide a collection of ethics-fostering interventions together with their strengths and limitations based on the qualitative data and interviewees' justifications.
3. We propose best-practice recommendations for improving ethics-fostering interventions and outline a path forward for community-driven development of ethical practices through an open wiki [89].

7.2 Methods

Answering our research questions requires a combined analysis of two qualitative data sets: (1) all publicly available calls for papers from top-tier S&P conferences necessary for addressing RQ2(a); and (2) semi-structured interviews with senior and junior members of the S&P research community necessary for addressing RQ1, RQ2(b), and RQ3. The following sub-sections outline our procedures for collecting and analyzing these two data sets.

7.2.1 Process-Data Analysis

7.2.1.1 Data Set and Analytical Procedure

A Call for Papers (CfP) states the topical goals of a conference and defines content-related and format expectations of conference organizers. CfPs can thus be considered historic documents which define key parameters for conference publications. Hence, it is a pivotal document setting key parameters for conference-based publications – the most precious currency within our community. Against this background, our general

Conference	Number of CfPs	Period
USENIX	28	1998 - 2025
Chairs' Message	12	2013 - 2025
NDSS	24	2001 - 2025
S&P	24	2002 - 2025
CCS	22	2004 - 2025

Table 7.1: Document corpus overview

RQ2(a) can be specified in the following way: How had the expectation of adherence to ethical norms been communicated in CfPs of the top-tier conferences, and what specific ethics interventions have been stipulated in these documents?

Against this backdrop, we analyzed all publicly accessible CfPs and submission guidelines of the four top-tier S&P conferences: CCS, IEEE S&P, NDSS, and USENIX Security (see Table 7.1). For USENIX, we also included the annual messages from the PC Co-Chairs. Two researchers analyzed all documents (CfPs $n = 99$; Chair messages $n = 12$) employing qualitative content analysis [207, 147] by coding each occurrence of ethics-related content (Holsti Index [144] of 96.9%) and clarifying disagreements. Thus, they identified various forms of ethics interventions.

In addition to publicly available materials, we analyzed review forms ($n = 35$) dating back to 2019. We collected them by asking senior community members to share documents from conferences for which they served on the PC. The lead researcher used qualitative content analysis to compile a list of ethics interventions and instruments.

7.2.1.2 Process Analysis Limitations

Our analysis primarily focuses on the S&P community. Although we surveyed a range of related fields, other scientific communities may have developed effective or innovative interventions that we were not able to identify. Thus, our proposed interventions may be only a subset of possible options. Nevertheless, we consider this selection a valuable starting point for a broader discussion about additional ethics interventions we want to deploy.

7.2.2 Interviews

Process data is an excellent source of information about official stipulations and procedures (e.g., the introduction of a REC). But they provide no insights into informal, often orally performed, activities or about the motivations behind certain publicly visible decisions. Since we are also interested in goals of ethics interventions, as perceived by our community's members, and in these members' assessments of existing and possible new interventions, we also conducted semi-structured interviews [75, 199], drawing on preliminary findings from our process-data analysis. These interviews can be characterized as expert interviews [26], because our aim was to tap into first-hand experiences of our participants with various aspects of ethics-fostering practices and their changes over time.

7.2.2.1 Interview Guide

We developed an interview guide drawing on the list of ethics interventions identified in the process-data analysis. It consisted of seven thematic blocks, each containing general and, where appropriate, follow-up questions addressing the following topics: (1) *ethics practices and reviewing procedures*, (2) *existing ethics interventions*, (3) *proposed ethics interventions*, (4) *ethics instruments*, the (5) *reasoning* and the (6) *goal* behind ethics practices, and the (7) *importance* of ethics. For the proposed ethics interventions and instruments, we made use of a survey instrument (Chapter 7.2.2.2) embedded into our overall interview guide. We also invited participants to share any additional thoughts. The complete base interview guide is provided in Appendix D.3. We adapted it slightly for different categories of research participants.

We conducted five pilot interviews to test our interview guide: three with former PC chairs and two with doctoral researchers. As the pilot interviews went without any problems, we kept the guide. We, however, added an explicit questions block about existing ethics interventions after the first two regular interviews; earlier, this was covered only implicitly via questions about ethics procedures.

7.2.2.2 Survey Instrument

Research ethics and corresponding interventions are very abstract topics. Thus, we expected less topic-aware participants to struggle articulating their views and reasoning. Hence, we incorporated a practical survey instrument into the interview to encourage more concrete statements.

Based on the process-data analysis, we identified potential improvements of ethics interventions beyond those already in place: (1) for *research ethics committees* to *consolidate RECs of different conferences into one inter-conference body*, *include external experts*, and implement a *REC-member mentoring system*; (2) *feedback to the authors* through *ethics meta reviews* and a *"profound ethical considerations" badge*; and (3) *education* with *educational material in form of categories* and *ethical training for reviewers*. This list of new interventions is not meant to be complete, but instead to gather first insights into interviewees' reactions to such innovations.

The survey instrument consisted of two pages (example in Figure 7.1 and all items in Appendix D.3). The first page presented the proposed interventions sequentially and in random order, and asked to rate each item using a five-point Likert scale (strongly agree to strongly disagree). The second page listed items for review forms: *ethics scores*, *feedback to authors* fields, and *internal discussion* options. Those were also rated on a five-point Likert scale.

7.2.2.3 Sampling

We specifically target members of the S&P community. We deliberately selected participants differing in their professional seniority: (1) very experienced community members who have served as PC or REC chairs and thus played central roles in shaping conference policies; (2) junior researchers, defined as paper authors without prior chairing

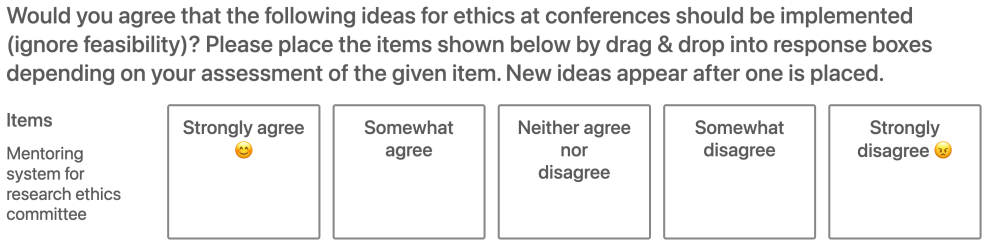


Figure 7.1: Survey instrument used as part of the interviews

experience. To ensure a diversity of perspectives, we further considered research background, gender, and nationality in our sampling procedure.

7.2.2.4 Recruitment

We used two different recruitment strategies. As for the senior members, we invited, on the one hand, individuals identified in our process-data analysis as directly involved in implementing significant procedural or structural changes across the top-tier S&P conferences, on the other hand, to cover a broad spectrum of opinions, we also included individuals who kept ethics procedures unchanged when they were chairing a conference.

For the junior members, we first determined a list of research fields to assign to them. To do so, we went through all available HotCRP instances of the major conferences from 2020 to 2024 and queried the unauthenticated `/api/searchcompletion` endpoint to retrieve the complete list of topics available at each conference. Based on this list, we manually merged semantically similar topics and removed redundancies to compile a holistic list of research fields. While this list (Table 7.3) is certainly not perfect, we found it to provide a sufficiently comprehensive and appropriately abstracted representation of the research landscape for the purposes of our analysis.

We then examined all papers published at the top-tier conferences in 2025 and assigned them to the research fields. We then randomly selected one paper from each field and invited the first author as a representative junior researcher. In cases where the first author was not a doctoral candidate, we invited the first doctoral candidate listed among the authors. This approach allowed us to recruit a heterogeneous group of early-career researchers across various research areas. To minimize unsolicited communication, we invited these authors by e-mail in batches of ten. Per author, we sent only one reminder after roughly a week if they did not respond. We subsequently proceeded with the next batch until we reached the coverage of nearly all research fields.

As a result (see Table 7.2), we interviewed 10 very senior members of the S&P community, among them multiple PC chairs and REC chairs, and 10 junior members who are mostly PhD candidates (with one having served on a PC). As the pool of potential interviewees is very small, we decided not to disclose further details on individual participants (and no institute location for senior people) to preserve their anonymity. Yet, we recruited interviewees from various regions, including North and South America, Europe, and Asia. We also included a wide range of research areas, covering all of our sampling matrix with the exception of *Wireless Security* (see Table 7.2).

ID	Seniority	Service Roles	Institute Location	Length	Ethics Training	Research Areas ⋆
I-01	Senior	PC-Chair	-	01:12		o
I-02	Senior	PC-Chair, REC-Chair	-	00:56		b, f, g, l, q
I-03	Senior	REC-Chair	-	01:28	Yes	a
I-04	Junior	PC-Member	US	01:05		c, h, f, j, k, l, n
I-05	Junior		NL	01:26	Yes	a
I-06	Junior		PT	01:29	Yes	l, o, t
I-07	Junior		CN	00:55		o
I-08	Senior	PC-Chair	-	01:36 (♣)	Yes	l, n, p, s
I-09	Senior	PC-Chair	-	01:06		a, g
I-10	Junior		US	00:42		l, n, t
I-11	Junior		DE	01:00		h, m
I-12	Senior	PC-Chair	-	01:09		e, h, f, j, k, l, n, u
I-13	Senior	Track-Chair	-	01:50 (♣)	Yes	a, j, t
I-14	Junior		DE	01:08		j
I-15	Junior		US	01:09		h, n
I-16	Senior	PC-Chair, REC-Chair	-	01:10	Yes	e, j, k, l, m, n, q, s, t, u
I-17	Junior		US	01:01		n
I-18	Senior	PC-Chair	-	01:03	Yes	i, p, t
I-19	Senior	PC-Chair	-	01:27 (♣)		a, h, i, l, p, t
I-20	Junior		AT	00:52		d, l, n
	Average	-	-	01:11		

Table 7.2: Interviewee backgrounds. ⋆ The letters correspond to research areas below. ♣ These interviews were split in two session.

a	Applied Cryptography and Cryptographic Analysis	l	Mobile Systems Security
b	Authentication, Access Control, and Authorization	m	Network and Protocol Security
c	Blockchain and Distributed Systems	n	Operating Systems and Software Security
d	Cloud Computing Security	o	Privacy and Anonymity
e	Cyber Attack (e.g., APTs, Botnets, DDoS) Research	p	Research on Surveillance, Censorship and Other Social Issues
f	Cyber-Crime Defense, Forensics and Diagnostics for Security	q	Secure Computer Architectures
g	Formal Methods and Language-Based Security	r	System Security
h	Hardware, Cyber Physical, Embedded and IoT Systems Security	s	Trustworthy Computing
i	Law, Policy, and Ethics of Security	t	Usable Security and Privacy
j	Machine Learning and Security	u	Web Security and Privacy
k	Malware Research	v	Wireless Security

Table 7.3: Research topics

7.2.2.5 Interviewing Process

Before the interview, every participant filled out a questionnaire with the informed-consent form about their research background and conference services.

Interviews lasted between 42 and 110 minutes. They were conducted between late April 2025 and early January 2026. We conducted and recorded all interviews via Zoom. After transcribing, we deleted the video file and only kept the audio for better coding flows. For transcribing, we used a local tool based on OpenAI's Whisper. A student assistant proofread and anonymized all transcripts.

7.2.2.6 Data Analysis

We conducted a qualitative content analysis [207, 147] of our interview data in a flexible manner: on the one hand, we used a pre-defined set of deductive codes derived from the literature on ethics in S&P research and our interview guide, but on the other hand, we were open to new data-driven insights captured as inductive codes created in a bottom-up manner (codebook in the supplementary material [88]).

Our research team consists of two cybersecurity researchers and two social scientists. At the initial stage, the lead author (cybersecurity researcher) and one of the social scientists coded five interviews (using ATLAS.ti CAQDAS package [10]) and discussed the resulting codings within the team. In this process, we realized that both coders had coded very similarly with pre-existing deductive codes. They reached a high inter-coder agreement: Holsti Index [144] of 93.3%. As for newly proposed inductive codes, both researchers discussed them until they resolved all divergences and reached an inter-subjective agreement on their use [149, 252].

After compiling an integrated codebook, the lead author continued with coding of the entire interview dataset. At the interpretation stage, the same two team members both extracted relevant information from transcripts using the shared code system and entered this information into tables which then allowed them to compare individual interviews along key categories (e.g., understanding of goals, assessment of specific interventions etc.). Next, they identified and systematized commonalities as well as differences in statements made by different participants. The final result was a systematic description of the spectrum of opinions articulated by interviewees.

7.2.2.7 Interview Limitations

Since it is a qualitative interview study, the frequency of opinions or judgments expressed by our participants must not be interpreted as indicative of frequency distribution within the entire population of S&P researchers. Our interview dataset can also be affected by the self-selection bias, skewing the interviewee sample towards those who are more ethics-sensitive and hence more willing to participate in such a study. Another limitation results from the *sampling bias* since we targeted popular S&P conferences and authors of published papers. As a consequence, we missed researchers who never published there - some of them possibly due to ethical reasons. We also assume that a study about a sensitive topic like research ethics could be influenced by the *social desirability bias*, i.e., participants may provide responses they perceive as more

socially acceptable rather than reflecting their true beliefs. Furthermore, we told our interviewees to ignore feasibility which probably made them more open for additional interventions, because resource considerations were faded out.

7.2.3 Ethical Considerations

We followed the ethical practices established in the S&P research community. The present study was approved by our university's Ethical Review Board. Before interviews, all participants were sent the consent form informing them about the content of the interview and their rights as participants. They all attended the interview voluntarily and were explicitly asked before the interview if they had consent-related questions or concerns. They had the option to discontinue their participation in our study at any time and to ask for deletion of their data. Besides these common interview-related ethical practices, we had two main challenges in mind that we considered when planning the present study:

How can participant anonymity be ensured when the number of potential participants is small? Since the participant pool of PC chairs is very small, it would be quite easy to de-anonymize individuals if not taken extra care about this aspect. We therefore decided not to publish demographic data about individual interviewees and also abstained from indicating the interviewee number for verbatim quotations and paraphrased transcript references. Furthermore, we have slightly modified some direct quotations to preserve their meaning while preventing the identification of interviewees.

How can participant recruitment within the community be conducted at a sufficient scale while minimizing unsolicited bulk mails, but also not conflicting with too many program-committee members? Another challenge was the recruitment of participants within the S&P community. A broad recruitment strategy via mass emails was not feasible, as it could have interfered with the double-blind review process. At the same time, repeatedly contacting and reminding a small group of individuals until they agree to participate would have been ethically unacceptable.

To get more perspectives on this problem, we first talked to former PC chairs about how they would handle papers that would conflict with too many PC members. Based on this input, we then decided to adopt a targeted and staged recruitment strategy (as described in Chapter 7.2). Senior members of the community were hand-picked based on their service track-record and contacted via targeted personalized e-mails. Junior members were randomly sampled based on their research backgrounds and contacted then in small batches. In both cases, we sent at most one reminder after approximately one week whenever no earlier response arrived. If no response at all was received, we moved on to a new set of candidates until our sample plan was met. During submission, we asked all participants who were part of the PC whether they would be fine with being marked as *other conflict* in HotCRP, explaining to them that it would potentially leak the fact of their participation to the PC chairs. As an alternative, we proposed that they can self-conflict with our paper. All affected participants were fine with the first option. We believe, in conclusion, this process holds a reasonable balance between minimizing unsolicited contact, avoiding conflicts of interest, and ensuring sufficient sample diversity.

7.3 Ethics at the top-tier Conferences

One of our interviewees pointed out that in S&P research, "*ethics has always been important, but not everyone understood, not everyone believed it was important or knew that it was important.*" As of today, ethical considerations have found their way into our community and its major conferences. So how has this happened?

In the following, we present a brief history of ethics debates within the S&P research community and the more recent introduction of various ethics interventions. We combine findings from three different sources: (a) the qualitative content analysis of process data; (b) our semi-structured interviews, especially those with senior members; (c) a quantitative keyword analysis of all papers published at the top-tier conferences (described in Appendix D.2). The CfP and the keyword analysis provide the factual skeleton of our presentation, while interviews capture informal processes and likely motivational structures behind the factual changes.

The keyword analysis demonstrates a significant increase of ethics-related vocabulary since the early 2010s (see Appendix D.2). While largely absent well into the late 2000s, by 2024[1] ethics-related keywords appeared in 22.3% of all papers accepted at top-tier conferences (CCS 16.5%, IEEE S&P 23.0%, NDSS 20.7%, and USENIX 28.3%). We thus ask what processes within the S&P research community led to this significant change.

Our interviewees were unanimous in describing the trend in the ethics debates: it was a movement from informal discussions, initially held at low-profile venues like co-located workshops (e.g., [152, 62]), towards an institutionalization of ethical reviewing and formal procedures (stated by 7 interviewees). The outcomes were: (a) ethical requirements and guidelines in CfPs, (b) formalized ethical-reviewing procedures, either conducted by a conference-wide REC, as it is currently the case at NDSS, IEEE S&P, and USENIX Security, or by track chairs within the domain of their scientific expertise, as at CCS, and (c) mandatory ethics sections introduced so far only by USENIX Security since 2025.

Interviewees recalled that the first very informal debates about ethical aspects started already in the 2000s. They were sparked by problems arising from research on botnets and honeypots and studies using leaked personal data from "*big data dumps*". At this time, major conferences were small enough for program committees to meet in-person in "*a conference room*" and discuss every submission, during which ethical concerns were raised before accept-or-reject decisions were made. Two eye-witness interviewees reported that ethics was back then "*an implicit question for the reviewers*" and PC chairs were able to "*keep an eye on this thing*".

Dedicated workshops on research ethics organized by the USENIX Association around 2010 (e.g., [152]) were described by one interviewee as "*the kick-off for the Menlo Report*" – the first milestone in institutionalizing ethics on ICT-related topics. Published in 2012, the Menlo Report was the first attempt to systematize, explicate, and promulgate an ethical framework for the entire ICT and S&P research. It outlines

[1]We omit 2025 since USENIX Security started requiring ethics discussions in all papers in that year, whereas writing about ethics depended solely on the motivation of the authors to do so, and probably also on requests made by reviewers.

some general ethical principles, but interviewees perceived it as not very "*actionable*" and too focused on specific problems of the time in which it was compiled. Our CfP content analysis indicates something remarkable in this context: program committees started to mention the Menlo Report only in 2022 (IEEE S&P), although the general topic of research ethics occurred in CfPs for the first time already in 2013. By 2025, NDSS, IEEE S&P, and USENIX referred to the Menlo Report in their CfPs. This fact suggests that the debates leading to the publication of Menlo Report were still confined to a smaller circle of ICT and S&P researchers with a particular interest in ethics, and had not yet reached the broader S&P community.

The first institutionalization step at the conference level was made by USENIX Security in 2013: a one-sentence reference to research ethics was included for the first time in the call for papers (process-data analysis) . NDSS followed in 2015. IEEE S&P and CCS mentioned ethics in their 2017 CfPs for the first time.

In the early 2020s, the top-tier conferences started to set up RECs, initiated by IEEE S&P and USENIX Security in 2022. Interviewees linked this trend partly to the public controversy after the acceptance of the "*Hypocrite Commits*" paper [95] at IEEE S&P 2021. The first REC at USENIX Security 2022 was still semi-formal: 15 experienced PC members commanding scientific expertise from different research fields were asked by PC chairs to deal with potential ethical issues indicated by reviewers [37]. In the following years, the USENIX REC became more formalized. As of early 2026, IEEE S&P, USENIX and NDSS have a REC, while there are, as of today, no signs that the CCS organizers plan to introduce a centralized REC for the 2026 edition and rather stick with its more decentralized ethical-reviewing procedure involving track chairs in pivotal roles.

According to REC chairs we talked to, the REC typically operates as follows: first, reviewers flag some papers as requiring additional ethical considerations in their review form. Then one or two REC members inspect the paper and assess handling of ethical issues, ask questions to the authors via HotCRP, and try to clarify open ethical problems. RECs have no formal power to reject a paper, but provide recommendations, with final decisions made by the PC chairs. In some cases, those can decide to publish a paper with a statement on ethics attached [36, 220].

Our interview dataset helps explain both key aspects of the trend described above: (1) Institutionalization: why research ethics have become so important for the S&P community that ethics-fostering interventions were integrated into the reviewing process. (2) Formalization: why reviewing procedures, including their ethics-fostering components, have become more formalized over time.

Regarding institutionalization of research ethics, our interviewees pointed to the growing ethical awareness within the community, the absence of IRBs/ERBs in many countries and an expanding presence of researchers from areas in which ethics traditionally play a significant role. The reasons for the increased sensitivity for ethical issues were mostly seen in some controversial publications, in particular, the "*Hypocrite Commits*" paper was mentioned most frequently (9 interviewees). However, one senior participant argued that this highly controversial paper had merely been "*a catalyst*" of further institutionalization steps rather than the main reason. Instead, they pointed to the general rise of ethical sensitivity within the entire society, impacting the ethical

awareness of the S&P community in the years following the COVID pandemic and the BLM movement in the US.

As for the formalization of reviewing procedures such as RECs, our interviewees pointed in particular to the numeric growth of the conferences in terms of submissions (more than doubling since 2020 [184]) and reviewers (5 interviewees). This scale made the formalization of reviewing procedures, including their ethics-related aspects, seem necessary to handle the exponentially growth of papers. The broadening of the topical scope of conferences and their increasing socio-cultural diversity were also sometimes mentioned as a reason for formalization which allows for a higher transparency of procedures necessary when community members differ in their understandings of research ethics.

7.4 Goals of Ethics Interventions

The assessment of ethics interventions, present as well as potential/future, requires an estimation of goals supposed to be achieved by these interventions as part of the overall reviewing practice. Hence, we examined our interviewees' explicit statements on goals of ethics interventions and the progress towards them.

7.4.1 Perceived Goals

We asked all our interviewees what they perceive as the main goals of ethics interventions introduced at the major S&P conferences in recent years. Their answers focused on four major issues: (1) normative regulation of various implications that S&P research activities (can) have for the outside world (on individual society members, certain social groups or organizations, society at large, etc.); (2) impact on the S&P research community itself; (3) influence on its individual members; (4) perception and evaluation of that community by various external actors. The second and the third issue are closely interrelated; hence, we discuss them in combination. Nearly all of our participants named more than one specific goal scattered across several of the above-mentioned categories.

The normative regulation of research activities w.r.t. their impact on the outside world was predominantly addressed by our interviewees as *avoiding harm*. In a similar vein, one participant stated that the goal was to prevent *malicious behavior* of researchers. Yet another interviewee defined the goal of ethics interventions as *protecting people* and *communities*. The same participant also mentioned environmental and computational resources, as well as *infrastructures* as aspects that should also be taken into consideration when assessing potential damage inflicted by S&P research.

Our research participants understand *harm* in a two-fold way: (a) a negative consequence directly connected to research activities (e.g., exposing human subjects to excessive stress during an experiment); (b) a negative consequence resulting from the publication of research results (e.g., a study disclosing an unknown vulnerability).

The impact of ethics interventions on the S&P community itself was most commonly framed as rising ethical awareness within it making authors think about ethical aspects of their scientific work and "*educating*" them. One interviewee even spoke of some

ethics interventions as "*a pedagogical exercise in terms of helping people learn how to do research in more ethical ways*". These interconnected aspects were most frequently referred to in the context of mandatory ethics sections introduced at USENIX '25.

Complementary framings focused on the impact of ethics interventions on the S&P community, emphasizing that it helped developing common community norms, making them more explicit and increasing the coherence of ethics-related decisions. One interviewee mentioned that institutionalized ethical reviewing can help to integrate a growing and diversifying S&P research community around core values. Another one added that it produces documentations which will allow to trace how ethical understanding and awareness of the S&P community evolves over time.

Furthermore, enforcement of ethical norms within the S&P community by denying publication of papers presenting research deemed as unethical was mentioned as an important goal specifically linked to the *right to reject*.

As already mentioned, some interviewees pointed out goals primarily focused on individual community members rather than the entire community. They emphasized that formalized ethical reviewing provides guidance for authors and reviewers and can also contribute to overall paper quality. We see these individual-level goals very closely intertwined with community-centered goals, because behavioral patterns of authors and reviewers shape major community-wide practices. At the same time, established community norms influence the behavior of individual members: those who are not able or willing to adhere to present ethical norms are sanctioned because they are denied publication opportunities.

The fourth type of goal proposed by our interviewees focused on how the S&P research community is perceived and evaluated by external actors such as private business companies, state agencies, legal systems, and the broad public. Ethics interventions were seen as ways to aim at improving the reputation of S&P conferences, research institutions and the entire academic science. In some extreme cases, these measures could also serve the purpose of protecting conference organizers from legal liability.

7.4.2 Progress Towards Perceived Goals

We also asked our interviewees to assess how far the S&P research community has progressed in achieving the above-mentioned goals. A clear majority of our participants stated that the implementation of ethics interventions is an "*ongoing process*". Some added that this process can never end, because its "*an iterative improvement*". There will always be something ethically incorrect happening and a need for the community to learn will persist. Nevertheless, the same research participants also admitted that recent ethics-fostering activities were moving in the right direction.

One interviewee presented a more ambivalent progress assessment: on the one hand, ethics interventions have already been fairly institutionalized, since "*mechanisms to check the ethics problems*" are in place. But on the other hand, there were still papers that got published although they "*have some ethics problems*". Two other participants went even further and stated that the goals of establishing ethics procedures at major S&P conferences have already been reached. A senior researcher argued: there are IRBs, ethical guidelines, and RECs at conferences; hence, no additional institutions

are needed. Many community members already do consider ethical aspects of their scientific work. As for ethically questionable studies that were possible "*15 years ago, we can't do them today*". In conclusion, this interviewee stated about research ethics: "*it's not as big as a concern*". A junior researcher pointed to the mandatory ethics section required by USENIX Security 2025 as the only substantiation of his claim that ethics goals have been already achieved.

Our study uses qualitative methods, and these do not allow to draw conclusions about the frequency distribution of particular opinions within the entire population of S&P researchers. Hence, we cannot conclude that a clear majority of the community perceives ethics institutionalization as an ongoing process, while only a few consider it as already accomplished. The views presented by two interviewees before offer valuable insights into perspectives skeptical of further institutionalization beyond the status quo. Notably, the goal perceptions of both participants were focused on external impact, specifically harm avoidance, and on protection from legal liability. They were the only two who did not mention community-centered goals, the otherwise most widely discussed set of goals among our other 18 research participants.

Key Takeaways

We identified 4 goal sets: (1) Impact on the outside world society; (2) impact on the S&P community; (3) influence on individual members; (4) external perception.

7.5 Ethics-Fostering Interventions

In the previous sections, we described the evolution of ethics interventions at S&P conferences and their goals as perceived by our interviewees. With these findings in mind, we now turn to interviewees' attitudes towards existing interventions, new proposals, and suggested improvements that came up during the interviews.

7.5.1 Existing Interventions

Over the past years, conferences have experimented with a range of ethics interventions. To get a more empirically grounded understanding of their effect on the community, we discuss the most widely adopted interventions below.

7.5.1.1 Right to Reject

Conferences reserve the "*right to reject a submission if insufficient evidence was presented that ethical and legal concerns were appropriately addressed*". It is the ultima ratio when no other intervention was successful.

Interviewees commonly agreed that conferences' right to reject is essential for sanctioning unethical behavior and enforcing community standards, echoing prior findings [188]. Without rejecting papers, some authors might not care about the community's moral standards, because "*if no one is ever rejected for unethical behavior, then there is no incentive for people who might want to act unethically*". One Interviewee also emphasized that rejection can serve as a learning experience, helping researchers to

reflect on ethical shortcomings of their research design: "*You know people got rejected [...] it's a learning experience. It doesn't mean you're a bad person. It just means you didn't think about certain things. And then the community is telling you, you shouldn't have done this.*"

However, participants also emphasized that rejection should generally be a last resort "*after a discussion with the authors and giving them an opportunity to comment.*" Ethical concerns were described as very case-specific, depending on the severity and irreversibility of harm: while issues like missing coordinated vulnerability disclosure can often be resolved in rebuttal, harms such as bad participant treatment cannot be fixed post hoc. One interviewee also supported rejection for formality issues like disclosure "*if it's explicitly stated in the call for papers, [as] it was the responsibility of the authors*".

As rejection is often case-specific, we also discussed how to handle papers with highly promising findings achieved with unethical methods. One participant argued against rejecting such work, suggesting that it would enable the research community to develop a follow-up solution, because "*a lot of brains working together and (...) if that paper got rejected then it's his (authors) own brain to trying to come up with a mitigation strategy*". Nonetheless, the general view throughout the interviews was that even highly impactful results should be rejected if obtained through unethical practices.

In the past, some conferences have accepted ethically controversial work but attached a public statement on the first page (e.g., [220, 36]). Interviewees expressed mixed views on this practice. On the one side, one participant said, the intent was "*to make transparent indeed what was this process and how the decision came to be*". Another added, it also deals as "*a warning that you don't do work like this*". On the other hand, such a "*paper has a bit of a stigma attached to it anyway, as it says right on the first page, which also has a slightly discouraging effect.*" It "*doesn't set a good example*", it would set a precedent, especially since preprints are often shared on "*open public servers, either ePrint or arXiv*", without the statement and such warnings are thus overlooked.

The fact that papers may often be published somewhere else is also a counterargument often raised when discussing ethically motivated rejections. Interviewees mainly dismissed this concern, noting that such publications have "*basically no impact*" and do not count in the academic "*bean counting*". One participant hoped, "*if the top conferences do that, will Tier 2 conferences eventually do the same, and will it be a kind of trickle-down effect*".

Key Takeaways

The right to reject papers on ethical grounds is widely seen as essential: it prevents the publication of unethical research (goal 1), maintains the conference's reputation (goal 4), and enforces community standards. It is viewed as the last resort when dialogue fails or harm is irreversible, even for highly innovative results. However, rejection alone raises ethics awareness only for the individual (goal 3) but does little to raise it across the broader community, highlighting the need for additional, more transparent interventions.

7.5.1.2 Research Ethics Committees

RECs are the operative bodies executing the in-depth ethical reviewing of papers flagged by reviewers. Consistent with prior work [188], our participants perceived RECs as helpful to "*validate, they review the concerns that people have. And they help trying to clarify these concerns*" with "*questions back to the authors*". Multiple participants emphasized that an explicit REC brings more structure to the reviewing process and helps keeping an overview of ethical aspects in our community.

Interviewees noted that RECs are a relatively new and still evolving body. Junior members in particular reported limited insight into how they operate and called for better transparency. Another complaint was that REC service takes time away from regular PC duties, an issue critical with the community expanding. Interviewees also mentioned that RECs are relatively small compared to full-PC discussions in earlier days, a scalability trade-off. While discussing ethical issues in the entire PC seems infeasible, one interviewee wished for "*more diversity in terms of backgrounds and geography*".

Some junior participants expressed the wish to consult RECs earlier about research ideas, though others views it as infeasible given limited resource capacity. However, according to some senior members, there was "*a service where you could potentially [go] if you had questions about ethics with your paper*", but it was discontinued due to little interest from the community and because "*cases that require an ethics committees inputs might not be ones where the authors [make in-depth] ethical consideration beforehand*".

Lastly, participants worried that not all papers are reviewed by the REC, because technically focused reviewers may overlook ethics. Hence, one participant suggested random REC sampling.

Key Takeaways

RECs are widely valued as a "second set of eyes" that helps to prevent publications of unethical work (goal 1), structure ethical procedures and raise awareness within the community through policies (goal 2), and provide ethics feedback to authors (goal 3). However, they should improve their transparency and diversity, and expand their scrutiny beyond flagged papers.

7.5.1.3 (Mandatory) Ethics Sections

A section to reflect on ethical considerations of a study has been standard in some fields like Usable Security, but USENIX '25 was the first top-tier conference to mandate them for all paper. Interviewees viewed these sections as necessary for some papers, but expressed mixed feelings on the mandatory requirement: "*It helps and doesn't help at the same time*".

More than half of the interviewees explicitly mentioned that ethics sections help them to reflect on normative aspects of their research. A key motivation for making them mandatory is seen in the fact that "*the papers where [...] we have the most ethical problems are the ones that don't realize that ethics is applicable.*" To support this change, ethics sections now get an extra page outside the page limit. It was seen as

necessary by interviewees because some projects require extensive ethical explanations.

Interviewees viewed the standardization as helpful for both PCs and authors. For the PCs, making researcher's previously implicit considerations explicit to readers helps resolve misunderstandings beforehand, and supports tracking ethical norm develop over the years. For authors, especially junior researchers, many participants noted ethics sections serve as guidance and also as inspiration for considering ethics in future studies. It was also emphasized that only little writing effort is required, and for papers with no ethical exposure a short statement explaining why no ethics is involved is sufficient.

Despite the low effort required, one interviewee noted about the new stipulation: "*almost anyone I've talked to hates it*", mainly due to the overly descriptive enforcement of the policy via desk rejections. Another participant warned that too strict rule enforcement could be counter-productive and may eventually lead to minimal-effort compliance (e.g., via LLMs), similar to how complex password policies led to password reuse and weaker passwords [101, 129]. Others countered that it is lastly the responsibility of authors to read CfPs.

Six interviewees noted that not all work required in-depth discussions of ethics, citing cryptographic studies, local code analysis, or SoKs as examples. Some researchers in these areas reported uncertainty about what to write and called for clearer guidance. However, USENIX counter-argued that as an application-focused conference, they expect ethics to play a role for every submission [246].

Hesitance around ethics and the lack of guidance for areas new to these considerations may have led to many vague ethics sections. Two interviewees mentioned they copy&paste from other papers, "*basically just repeat(ing) the same words again again again*" , or used LLMs to write this section, reflecting minimal-effort compliance. One senior interviewee viewed this as a transitional issue rather than a major concern, addressing it through review comments.

Lastly, two participants questioned the cost-benefit balance, arguing that historically, few submissions had raised serious ethical concerns. They warned not to punish everybody for the wrongdoing of few, especially in light of other challenges such as collusion rings [140], which would require more countermeasures.

To make ethics sections more structured and to reduce the risk of formal rejections, several participants suggested adopting checklist templates similar to those used by NeurIPS [169] or ICWSM [99]. Such lists should contain standardized questionnaire items as well as open-ended fields to keep flexibility.

Key Takeaways

Mandatory ethics sections encourage authors to reflect on the ethics of their research (goal 3) and, given the extra page, have little cost. Making them mandatory enforces this awareness across the entire community (goal 2). At the same time, limited guidance and overly strict enforcement risk vague compliance, resistance, and policy bypasses; a low-barrier checklist template was suggested.

7.5.1.4 Ethics Framework and Policies

Lastly, we covered several policy-related interventions: (1) the request to disclose IRB decisions in papers, (2) suggesting coordinated vulnerability disclosure guidelines, (3) using field- or method-specific ethics categories to prompt relevant considerations, and (4) recommending general ethics frameworks.

As for (1) IRBs, interviewees pointed out they "*are necessary, but not sufficient.*" There are many reasons for this assessement, as noticed in literature [77, 188]. Our interviews pointed out that IRBs are country- and institution-specific and often lack technical expertise. Still, two interviewees argued: if authors would submit IRB material alongside their submissions (similar to artifacts), it may reduce later REC requests. Our CfP analysis shows that mentioning IRBs in the paper was often demanded in the past, but shifted, e.g., CCS '26 states that IRB approval is not strictly required and ask authors rather to reason about ethics beyond IRB requirements. This conference seems to follow the recommendation by Ramulu et al. [188].

Interviewees agreed that (2) coordinated vulnerability disclosure is crucial and "*probably the one area where we [...] have the most experience, and therefore we have kind of the most potential rules*". Guidelines were seen as especially helpful for new people in the community, also since disclosure is often handled with support from senior group members. Nevertheless, some participants emphasized the need for flexibility citing various reasons for diverting from standards, e.g., previous bad experiences with specific vendors who broke embargos or hardware vendors with long product cycles.

Method-specific ethics categories (3) were viewed by interviewees as useful to prompt authors to very field-specific concerns, noting that the community is too broad for *one* set of rules. However, categories cannot cover all edge-cases and must be adopted with new developments, like AI research. For example, one theorist had difficulties mapping their work to a category and then turned to related work and LLMs for advice. Also, several participants called for example papers alongside categories, noting that "*it's good to have at least examples [...] as to what you need to discuss.*"

Regarding (4) general ethics frameworks (e.g., stakeholder analysis), participants have a similar opinion as before. They provide helpful guidance, particularly for authors writing an ethics section for the first time. One interviewee recalled their USENIX '25 experience: "*I thought more about ethics than I did for any other conference*" and now uses this reflection as a baseline for future papers. Stakeholder analysis was also positively assessed, for helping researchers to better "*understand their own research and where their research fits in [the] world better*", eventually improving the paper overall.

However, as before, participants emphasized the need for flexibility to divert from proposed frameworks. RECs could instead request more structured analysis during rebuttal if needed. They argued that ethical issues vary across research areas for which some frameworks might fit better than others. Also, the frameworks we use might not reflect the socio-cultural diversity in our community, as "*what happens when we impose an ethical framework that is based on a particular value system on researchers who may have a different one. Do we say, well, our way is best?*"

Lastly, our interviewees also criticized "*that, like every year for the same conference, they change the guideline.*", which makes it difficult for authors to follow. Another interviewee noted that it could increase desk rejects due to minor mismatches and add

burden of rewriting ethics sections for different conference frameworks. Many participants therefore emphasized: there should be community-wide consensus establishing stable and shared ethics standards.

Key Takeaways

Ethics frameworks and policies help to reach goals by outlining ethical expectations for harm-avoiding research (goal 1), shaping community norms (goal 2), guiding authors and reviewers by giving structure to ethical considerations (goal 3), and presenting the community's ethical norms to external people (goal 4). However, they should remain flexible in content, yet consistent in process across conferences.

7.5.2 Problems and Challenges

The analysis of historical developments, existing interventions and other interviewee statements allowed us to identify a number of challenges affecting ethical-review procedures.

(P1) Growing community: The growth of the S&P community increases workload for conference boards, amplifies reviewer fatigue, and consequently slows down PC discussions. This constraint hinders the community's ability to approach issues effectively.

(P2) Diversity: As our community gets more diverse, there is a need for the RECs to become more diverse and inclusive.

(P3) Education: The topic of ethics is often taken for granted, but some members of our community, especially those who joined recently, miss education and guidance on the shared norms.

(P4) Transparency: Transparency at both levels, conference-wide decisions (e.g., rule changes, REC procedures) and review level (e.g., feedback on ethics sections), could be improved especially for researchers not yet embedded in senior community circles.

(P5) Regulatory Balance: Conferences need to properly calibrate the scope of formal policies and their enforcement to avoid overregulation and unintended backlash.

(P6) Consistency: There was complaint about inconsistencies in conference procedures: "*[E]very PC chair wants to conduct a few experiments. [...] experiments are good, they help us learn more, but we conduct a little too many experiments without evaluating them.*"

(P7) REC after research execution: Interviewees also highlighted the well-known problem [58, 188] that RECs assess studies only post hoc, unable to prevent unethical studies in advance, a problem that we also discuss in Chapter 5.

7.5.3 Proposed New Interventions

After identifying four perceived goal sets and key implementation challenges, we now discuss interviewees' assessments (see also Figure 7.2) of additional ethics interventions which we propose with the aim of addressing the list of problems (*Ps*) summarized in Chapter 7.5.2.

7.5.3.1 Improvements to Research Ethics Committees

As a first group of items, we proposed improvements to the REC.

Consolidate all RECs into one unified board for all top-tier conferences. Besides unifying ethics reviews, the idea was to have one board that considers and discusses all mentioned *Ps*. Overall, our participants liked the idea (average Likert score (LS): 4.00), but saw difficulties in the practical implementation.

Participants emphasized that "*we submit the same papers to the same place [...] We have the same reviewers*", and "*research ethics is a common problem [...] to solve this problem, we need to be more efficient*". One participant even noted that they saw a paper rejected on ethical grounds at one top-tier conference but accepted at another. Therefore, they would appreciate a unified group that would introduce more consistency and efficiency to ethics reviews.

However, even supportive interviewees raised concerns regarding its feasibility. First, conferences have tight timelines and "*getting enough people to work across multiple conferences is a challenge*". But the larger concern is the institutional fit as "*PCs are maybe the same, but the management is quite different*", raising questions of funding and accountability. As a middle ground, interviewees suggested a inter-conference body in charge of general principles and policies, while leaving implementation and case handling to individual conferences.

Include external ethics experts into RECs to add perspectives beyond the S&P community. This interventions addresses *P2* and was strongly supported (LS: 4.60) with interviewees stressing that "*more voices, the better*" and viewing it as "*something [...] that brings less load to the people with a very big gain*". It would "*help existing members of the community understand practices from beyond and incorporate those into our own conception of ethics*".

Interviewees suggested various experts, including philosophers and sociologists, basically "*people who are actually concerned with ethics, who studied ethics*". Some also advocated to include legal experts to advise on data protection, regulatory compliance, and potential legal implications of studies.

Nevertheless, participants also raised some concerns, including potential community barriers (e.g., terminological differences), a lack of domain knowledge, and an increased coordination overhead that could result from communication with external experts.

Notably, interviewees highlighted that USENIX Security already integrated philosophers into its REC in 2025, pairing them with technical members for joint reviews, an approach reported as effective in bridging domain and communication gaps.

Mentoring system for RECs. Our idea was to pair a junior REC member with a senior one for a joint review during the first year, followed by independent reviews by the junior the following year, and the option to mentor a new member afterwards to address *P1* and *P2* in the third year.

This idea was positively received (LS: 4.25). Participants viewed it as a way to diversify the REC as "*more people, including junior staff, may have different perspectives on such matters*", and to standardize REC member recruitment, which is currently ad hoc: "*we chose people [...] because we know that they [...] have worked previously on some ethics on papers*". It is also seen as low-effort way to spread knowledge of ethical norms within our community.

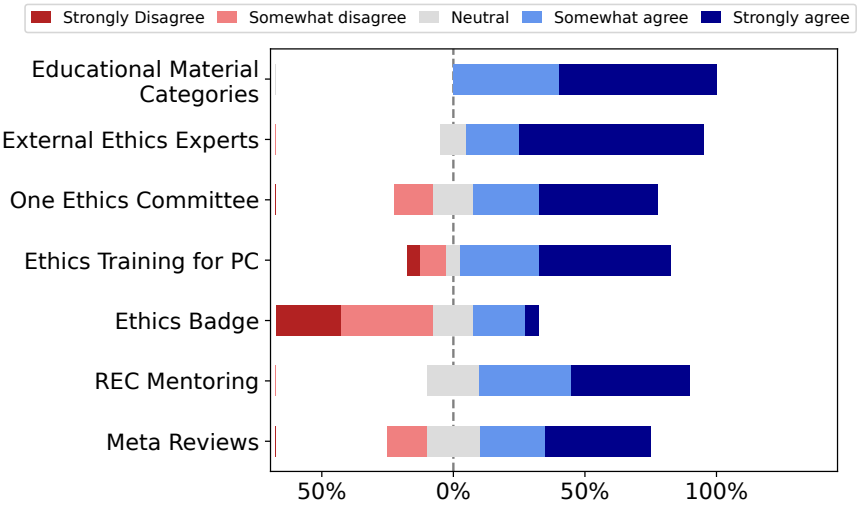


Figure 7.2: Likert scale of interviewee's opinions about proposed interventions

While some saw mentoring as low overhead, one raised a concern about the time costs required for good mentoring and about long-term sustainability with keeping the same members involved for multiple cycles. Other concerns included remarks that mentors would bias ethical judgments, and skepticism about the added value, as ethics is mainly shaped through life experience. One senior participant also felt uncomfortable mentoring without formal ethics expertise, but would support it if voluntary.

7.5.3.2 Feedback to the authors

Some participants, especially junior researchers, noted missing feedback (*P4*) on their ethical considerations, motivating our proposals for ethics-focused author feedback.

Ethics meta reviews. We proposed to use meta review as approach to summarize PC ethics discussion per paper. Interviewees viewed it positively, but less strongly than other ideas (LS: 3.90), with senior members (LS: 3.70) slightly less supportive than junior members (LS: 4.10).

Supporters of ethics meta reviews noted that more feedback and transparency is good. "*[D]efinitely people need some visibility of what was discussed or otherwise they don't learn*". This could help others see "*what was good about this paper, what was wrong with this paper*". For example, one author wondered about another paper, why it was accepted and would have liked to see its meta review by the PC or REC. One interviewee also noted there is no reason not "*to publish these reviews because actually it has already costed the reviewers some time to discuss this part*".

More skeptical interviewees, mostly senior, questioned whether the benefits justify the added workload given reviewer fatigue, and whether anyone would read meta reviews. Other senior members expressed skepticism regarding the content of such meta reviews: since ethics discussion are often only about asking authors whether they "*did or did not address their issue*", there is not much interesting to learn from. Also, these discussions may include sensitive details that should not be public, e.g., if "*researchers*

choose to knowingly violate the law in a minor way [...]. This is a challenging conversation to have, but I don't know if we want a written [statement]".

"Profound ethical considerations" badges. The idea of a badge was in analogy to artifacts evaluation badges dealing as feedback (*P4*) for authors and good examples for young researchers (*P3*). However, participants in our interviews were largely skeptical about this idea (LS: 2.45).

A few interviewees saw an ethics badge as a useful way to highlight exemplary ethical considerations for others to learn from and as good feedback "*to reward the one[s] that go beyond*" the baseline.

However, the effort put into the considerations are not necessarily reflected in the section itself and ethical complexity varies widely across areas, as multiple participants noted by raising fairness concerns. The main criticism was that, unlike artifacts evaluation, ethics is too subjective to be measured with clear criteria, leading to uncertainty about what the badge would actually stand for. Many also argued that ethical considerations constitute a baseline requirement for a paper, implying that every paper would get the badge.

Instead, some interviewees suggested alternative forms of recognition, such as community awards to honor distinguished ethics sections to inspire future researchers with good examples.

7.5.3.3 Education

Little ethical education (*P3*) beyond general university courses was a recurring concern and may hinder ethical awareness (goal 2 & 3) within our growing (*P1*) community. We therefore proposed several education-focused interventions.

Method-specific educational material. Building on USENIX Security's method-specific tips [16] (e.g., experiments with live systems; deception studies), we discussed method-specific guidance for researchers. Interviewees strongly supported it (LS: 4.60).

Interviewees described such material as necessary, especially for researchers unfamiliar with ethics: "*I think it is necessary to provide some materials to help them understand.*" One valued method-specific categories as concrete prompts as "*good. If you can give me categories, which can make it easier to think about ethical concerns*", and may also help to find some issues that one might not have thought about. Overall, participants stressed that S&P researchers are not ethics experts and need more education on this topic.

One junior participant noted that many authors would only consult such material superficially, admitting "*I would not go through the education material before the ethic section.*" This raises the question if the effort creating such material is worth the benefit, given limited resources in the community. Interviewees also raised concerns that ethics is subjective and method-specific categories might be hard to define optimally. Thus one participant would rather like to see good example case studies over fixed categories.

Multiple interviewees expressed the need for concrete example studies – rather than abstract scenarios – as those would ease considering ethics. One of them suggested a database of real studies linked to specific ethical policies, and noted mandatory ethics sections now already create the raw material for this as we get "*more data on what*

people have thought about, which you can possibly draw on if you find yourself in a similar situation".

Ethical training for reviewers. The idea of this intervention is to address the mentioned problems (*P1 & P3*), and the consistency of decisions (*P6*). This suggestion was received positively but raised concerns about time constraints (LS: 4.10).

Training was seen as a good option to familiarize reviewers with conference-specific ethics policies as some PC members seem to be unfamiliar with them, and to establish a shared baseline, given that most S&P researchers lack formal ethics expertise.

One interviewee, however, noted that reviewers nowadays need ethical expertise mainly in their own subfield with which they should already be familiar. Another participant was afraid that training could become overly rigid and "*brainwash*" trainees to adhere to one specific ethical approach, constraining reviewers' independent judgment; teaching general ethical ideas was seen as acceptable. Others asked who would provide the training, particularly against the background of different ethical traditions and socio-cultural backgrounds. And even with sufficient training, some interviewees wondered whether it would pay off, since "*reviewers who don't care about this will [...] just skip over this as fast as possible.*", while those who care are already knowledgeable about it. Hence, "*there's a difference between providing the training and ensuring that people take it*". Above all, the main concern was the severe time constraints and reviewer fatigue: "*we already scramble, we already struggle a lot to get enough reviewers [...] it's a mess.*"

Some interviewees pointed out that many researchers already receive ethics training from IRBs and suggested: "*if the reviewer has participated in the training [...] they should not need to participate in the same training for conference B in the same year*". Lightweight briefings on updated ethics policies were widely appreciated.

7.5.4 Review Form

As outlined before, reviewers serve as the primary agents of ethical scrutiny in the publication process. RECs become involved only when reviewers identify potential ethical concerns and flag them in the review forms. Consequently, the inclusion of well-designed ethics instruments within these templates represents a significant ethics intervention in its own. However, one senior interviewee stated, "*the USENIX Security 26 review form is too complicated. [...] that's feedback I've heard from various people.*" Against this backdrop, we discussed with our interviewees several approaches to communicating ethics-related considerations via review forms. We present the identified findings below and in Figure 7.3.

7.5.4.1 Ethics Rating

We discussed three ways to rate or flag ethical concerns in reviews: a **binary flag**, a **numeric scale**, or a **severity score** (e.g., *no issue*, *minor*, *severe*). Interviewees were divided on the use of scores, but agreed on two points: (1) numeric scores are the least useful (LS: 2.55) since they offer little added information beyond severity. (2) free-text justifications along scores is essential.

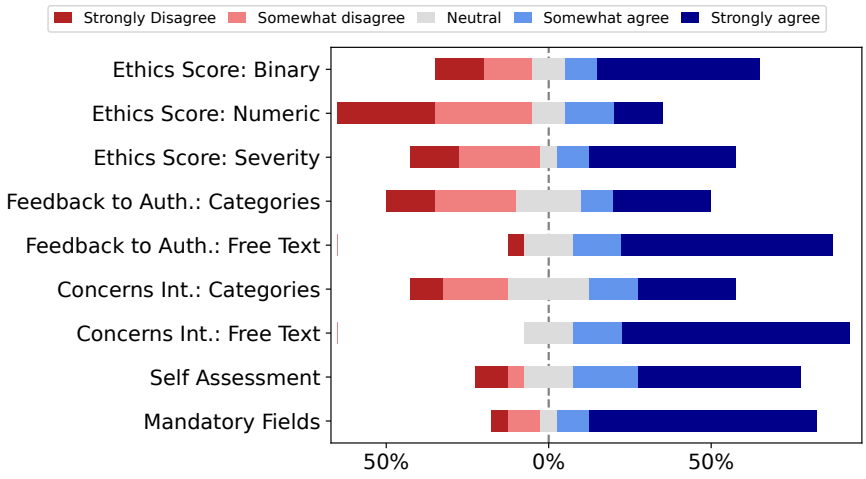


Figure 7.3: Likert scale of interviewees' opinions about ethics instruments in conferences' review forms

Severity score (LS: 3.45) was valued by some as a way "*to give [...] the chairs some guidance*" and help them to prioritize cases. Also, some noted, ethics is too nuanced for a mere yes or no. On the other hand, other preferred a binary flag (LS: 3.65), arguing a paper either raises ethics concerns or not. Three interviewees rejected scores altogether as they make reviewers think less about ethics, stressing that "*what is important is, that people tell that there is a potential issue and explain why and that then other people have a look at it*".

7.5.4.2 Ethics Concerns

Regarding communicating concerns, we discussed **PC-internal feedback** and **feedback to the authors** in form of either **free text** or as **pre-formulated categories**.

Across both channels, interviewees strongly preferred free text (LS: authors 4.35; internal 4.55). Categories (LS: authors 3.15; internal 3.35) were seen as a supportive instrument, but not sufficient on their own. Because some ethics problems occur frequently (e.g., missing information about a CVD process) a lot of ethical feedback is paper-specific and pre-formulated categories cannot cover all cases. Also, regarding feedback to authors, one interviewee emphasized that every rejected submission deserves that we "*make the effort to write a few sentences [...] explaining why we don't want this.*" For internal feedback, some senior members pointed out that it helps the REC to internally assign papers to REC members with corresponding expertise. Also, one noted that internal feedback is important, because sometimes, "*you'd want to raise to the committee that you wouldn't necessarily want to raise to the authors, because you're unclear whether something constitutes violations*"

7.5.4.3 Additional Proposals

Lastly, we proposed a **self-assessment** score for reviewers' ethics confidence and the requirement to make ethics fields in reviews **mandatory**.

The self-assessment was viewed as moderately useful (LS: 3.95), mostly for helping RECs to prioritize and double-check papers with low reviewer confidence. However, criticism was that it could lead to situations where reviewers or authors ignore valid concerns expressed by less confident members. One participant also noted that RECs should review every flagged paper regardless of a reviewer's self-assessment. Most interviewees supported mandatory ethics fields in reviews (LS: 4.30). Critics mainly argued that not all papers have ethical aspects, but were comfortable with mandatory binary flags or scores as long as they can indicate *no concern*.

7.6 Discussion

One can only reason about good or bad practices if knowing intended goals. We therefore identified the perceived goals of research ethics in the S&P community and use them as a baseline to discuss interventions, encouraging decision-makers to do the same.

7.6.1 Goals of Ethics Interventions

According to our interview analysis, members of the S&P research community perceive four sets of goals behind ethics interventions at top-tier conferences (Chapter 7.4). The examination of the existing interventions (Chapter 7.5.1) suggests that normative regulation of external impacts of S&P research (goal 1) as well as external perception and evaluation of the S&P research community (goal 4) have been already addressed by such institutional measures like ethical requirements written into CfPs, ethical review integrated into the routine reviewing process and supplemented by reviewing activities of specialized bodies like RECs, and ultimately by the right to reject publication of unethical work. This fact can prompt some community members to think that there is little left to be done with regard to the institutionalization of research ethics. However, many interviewees indicated that a further increase of ethical awareness within the S&P research community and support of ethically sound practices of its individual members constitute another major set of goals (2 and 3) supposed to be addressed by ethics interventions.

We see these four sets of goals as interconnected. Reputation and assessment of the S&P research community by external actors (goal 4) depends to a large extent on the ability to effectively regulate in normative (ethical) terms various impacts of research activities and results on the outside world (goal 1). And the approximation of the latter goal is significantly influenced by the community's ethical awareness and the ability of its individual members to put the principles of ethically sound science into practice. Against this background, we propose in the following subsections an expansion of ethics interventions specifically aiming at a further rise of ethical awareness within the S&P research community and making it easier for individual community members to integrate ethical reflection into all stages of their scientific work.

7.6.2 Ethics Interventions Recommendations

As we mentioned before, the community made strong progress on ethics interventions. Still, we identified a few improvements that we as a community should discuss, consider, and think about how to put considerations into actions.

Rejection is the ultima ratio and needs careful consideration. Rejection stands above all other interventions, which either inform or operationalize this decision. It is largely accepted and uncontroversial within the community. Disagreement arises only from ethically ambiguous research with great results, for which some have decided to publish with an ethical disclaimer. The impact of this approach remains unclear [188]. Irrespective of whether the work is accepted, the harm to the outside world (goal 1) has already occurred in such cases. Given the lack of a community-wide standard for ethics, rejection may lead to unethical work being published at another conference, causing a negative impact on the community (goal 2). Similarly, not publishing the work may lead to other researchers conducting the same or similar experiments, causing additional harm to the outside world. However, publishing with a note may suggest that transgressions of ethics are acceptable as long as the results are sufficiently interesting, leading to negative incentives (goal 3) for researchers and potential reputational damage to the conference (goal 4). More positively seen, publication with disclaimers sets a precedent for what should not have been allowed to be published, making it an explicit signal for future authors. We believe this requires community-wide discussions.

Research Ethics Committees can improve with greater diversity, more transparency, and better coverage. RECs already add great value to the peer-reviewing process by structuring ethical review procedures (goal 2) and providing feedback to authors (goal 3). However, we found concerns about limited transparency and diversity, and that unethical work may go unnoticed when reviewers focusing on technical aspects fail to flag it. More perspectives and greater coverage would help to reduce blind spots in identifying harmful work (goal 1), while increased transparency would strengthen the acceptance of norm enforcement within the community (goal 2). To improve, we suggest REC mentoring, including external experts, and, if full coverage is not feasible, at least checking random samples.

Ethical policies should guide practice without becoming over-restrictive. The community likes policies as guidance, and we found indicators that example policies (e.g., USENIX Security) are reused as a baseline for future work (goal 2). At the same time, we should be careful with overly restrictive rules, as people have called for flexibility. We otherwise risk negative effects and pushbacks (goal 2), such as minimal-effort compliance or policy circumventions (see complex password policies). We encourage decision-makers to open their policies to community input, as it was recently done by ACM CCS '26 [49].

Ethics section should be mandatory, with clearly defined exceptions. While mandatory ethics sections remain controversial as a restrictive intervention, they seem logical towards the awareness goals (2 & 3) and typically require little effort. While the community still discusses whether every paper needs one, we argue that a structured, uniform approach has benefits and thus recommend it for every paper. For example, the USENIX '26 Ethics Guidelines provide four key questions (stakeholders, impacts, mitigations, decision to proceed) that could serve as a lightweight checklist for

authors to self-certify whether they should be exempted from discussing ethical implications. This way, all researchers have to reflect and self-assess their ethical implications before submission.

The community needs shared, community-wide ethics policies and educational material. While CfPs provide valuable guidance [188], frequently changing, conference-specific rules lead to frustration in the community and potential increased reviewer fatigue. We thus recommend developing community-wide standards (goal 2 & 3) that can be adopted across the top-tier and other conferences (trickle-down effect). Beyond policies, shared resources should include research-area-specific educational material, concrete example papers, as researchers often learn community-specific ethical understandings from studying related work. Also we should discuss options to publish anonymized REC rejection decisions as a way to communicate community boundaries.

The community would benefit from a coordinated ethics steering body. Conferences established RECs to guide ethical decisions. While conferences need these to take on conference-specific decisions, we advocate for a community-wide ethics steering body to support the above-mentioned shared policies and coordination. Such a body could better scale with the growing community, address transparency concerns, advise on ethically complex research at the design stage, and help follow the identified goals.

7.7 Summary

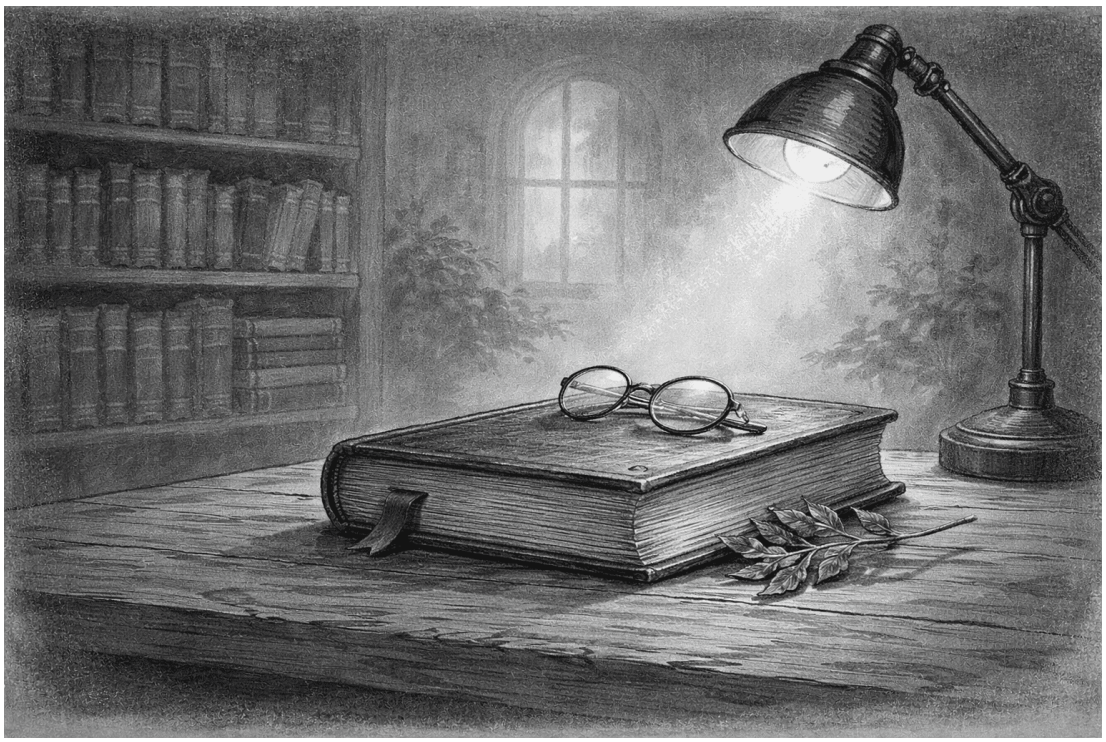
This chapter presents the first study on existing ethics policies and goals, challenges, and perceptions surrounding ethics procedures in top-tier conferences. Our findings suggest that while the community broadly agrees on key objectives such as raising ethical awareness and preventing the publication of unethical work, progress is hindered by limited ethics education, inconsistent policies across conferences, and concerns about over-regulation.

To address these challenges, we highlight the need for greater reflection and discussion about ethical practices in the community by publishing domain-specific ethics considerations and coordinating ethics efforts across conferences. A common understanding and agreement on ethics policies can help strengthen responsible research practices while avoiding unnecessary procedural burdens. We hope that our findings contribute to a more informed discussion on ethics policies and support future efforts toward transparent, effective, and community-driven ethics practices in security and privacy research. In turn, these interventions can help improve how security and privacy web measurements are designed and conducted.

Part III

Final Remarks

8
Conclusion



In this dissertation, we addressed the broader question, *"How can we perform reproducible and responsible web security and privacy measurements?"* We now conclude with a reflection on the key takeaways that have emerged from the presented studies, and discuss the remaining challenges, future research directions, and the broader picture to further improve web security and privacy measurement.

8.1 Reproducibility Through Advancement of Techniques

The Web is a constantly evolving system. Websites frequently change their content, structure, and behavior, sometimes within hours; and with AI, the pace of change is likely to increase further. As a result of these changes, reproducing web measurements on the live Web is challenging, if not impossible. However, reproducibility is a fundamental principle of scientific research, as it allows independent teams to verify results and enables researchers to build on prior work. Ensuring reproducibility is therefore particularly important for empirical web security and privacy measurement research.

Summary of Contributions. We investigated approaches to address this reproducibility challenge. In Chapter 3, we demonstrated and evaluated the potential of existing web archives as a reproducible database for S&P measurements. Our results show that the IA can already support certain measurements, especially those of static client-side data such as HTTP security headers. Because archived snapshots originate from different contributors and vantage points, they provide a diverse perspective on the Web. However, this diversity also introduces variability between individual snapshots. To obtain more representative results, we therefore propose analyzing not only individual snapshots but groups of temporally close snapshots, which we refer to as neighborhoods.

While existing web archives can support measurements on static content and even some dynamic content, they do not work for sites that rely on dynamically live exchanged content or live data from web APIs. In Chapter 4, we therefore explored how an archiving tool and format for web security and privacy measurements should be designed. We argued that reproducible web archives must capture not only page content but also page behavior during execution. To address this requirement, we proposed and evaluated WebREC, which is based on a technology that records browser activities in addition to network resources. This approach enables the creation of reproducible measurement artifacts that preserve the behavior of websites at the measurement time.

Challenges & Future Work. Despite these advances, several challenges remain. Currently, WebREC is implemented on top of Brave. Extending support of the *.web* archiving format to additional browsers would improve coverage and allow researchers to study browser-specific behavior. Furthermore, WebREC focuses on the page context and thus does not support experiments involving browser extensions or service workers. Extending and evaluating WebREC in this direction to capture emerging threats from extension malware could further broaden its applicability for reproducible web research.

Another challenge concerns the representativeness of archived measurements. Researchers must carefully consider what exactly their measurements aim to represent:

(1) Should a study capture trends from an individual perspective or observations from different vantage points across the world? (2) Similarly, should the focus be on the most popular (and probably most relevant) sites, or a broad view of the Web (including small, very outdated hobbyist sites)? These choices all make a difference and can lead to different conclusions. We argue that ideally, an experiment, as well as the archives, should include crawls with multiple configurations and vantage points to avoid presenting a one-sided view of the Web and represent sites most relevant to the population under study. Developing improved sampling algorithms and analysis techniques that account for such (1) variability and (2) coverage, similar to our concept of neighborhoods, would therefore be a promising direction for future work.

The Broader Perspective. From a broader perspective, the ideal scenario would be one in which researchers do not need to execute their own measurements at all, but instead can reuse comprehensive archives that contain the necessary data for a study. Consolidating the data collection process could reduce errors during experiments and, from an ethical viewpoint, minimize the need for repeated scans of the same systems by different research groups. Projects such as the HTTP Archive already offer this by providing large-scale web measurement datasets. Extending such datasets with behavioral information, such as the PG traces captured by WebREC, could immensely improve their value for security and privacy research. However, that is not enough, as modern websites often execute different dynamic code paths depending on user interaction such as clicks and scrolls, making it difficult to capture all relevant behavior automatically. Future research could therefore explore archiving systems that improve action coverage by analyzing code and triggering relevant interactions during measurement.

At the same time, this vision of a universal archive may be difficult to realize in practice as it would require an immense amount of resources. Moreover, while we showed *.web* already captures types of commonly used data for research, archives will likely never contain every specific piece of information researchers sometimes need. Consequently, the community should invest in technology for hosting and sharing these datasets, and even more so, in standardizing and advancing reproducible measurement tools and artifacts – such as the *.web* format – so that individual artifacts from research teams can together form a reusable and growing database of web measurement data.

In conclusion, enabling reproducible web security and privacy measurement research requires both improved datasets and improved methodologies – improved techniques. Archival approaches can already support reproducibility for certain types of measurements, while new tools and formats enable the capture of complex web behavior. Moving forward, the research community should strive to design measurement infrastructures and datasets with reproducibility as a default goal, thereby fostering more transparent and verifiable web security and privacy research.

8.2 Responsibility Through a Good Discourse

The Web has become a ubiquitous system. Websites are widely used by individuals and organizations to handle sensitive tasks and confidential data. As a result, poorly designed or incautious experiments can cause harm not only to technical systems but also

to the people and organizations relying on them. At the same time, public confidence in science and academic institutions has increasingly been questioned in recent years. In this context, responsible research practices are particularly important, especially for empirical web security and privacy measurement research.

Summary of Contributions. Across the second half of this dissertation, we have discussed ethical challenges in web security and privacy measurement research and practices at the academic conferences. In Chapter 5, we presented insights from interviews with legal experts, ethics committee members, and web operators, discussing questions, scenarios and legal and ethical boundaries around server-side scanning practices. Our findings highlight that many challenges arise from still existing legal uncertainty, particularly in an international web ecosystem where regulations differ across jurisdictions. At the same time, the study shows that we need more ethical awareness among researchers and discussions in all security domains about what is right and what is wrong. Importantly, the operators we interviewed were generally open to academic measurement research, provided that researchers act transparently and responsibly.

Because awareness and more ethical discussion were named as one key challenge for researchers, we looked in Chapter 7 at ethics practices and interventions at top-tier conferences. Our findings reveal four goal sets, challenges, and attitudes toward ethics interventions. While measures such as ethics requirements in CfPs and RECs address harm prevention, ethical awareness and education still need improvement. Interviewees called for clearer guidance, educational resources, and more transparent procedures, and expressed frustration with inconsistent policies across conferences and years.

Open Questions & Future Work. Despite the progress made in recent years, several open questions remain. The frustration with the inconsistency of conference policies raises the question, how should these policies eventually look like? In Chapter 7, we made suggestions based on a qualitative sample of people; however, it is necessary to understand, with a quantitative survey within the research community (not only Western views, but also international ones), how the community stands on this topic. It would help to further guide decisions of the conference heads and community-wide efforts to shape commonly accepted rules.

Moreover, our impression from the interviews is that researchers who actively engage with ethical questions are often those already interested in the topic. A key challenge, therefore, lies in reaching researchers who may not yet be aware that their work could have negative consequences. Collaboration with experts in education sciences could help develop better ways to integrate ethical reflection into research training.

One important realization from this dissertation is also that certain types of server-side measurement research may also simply not be feasible on the broader Web, either for legal or ethical reasons (e.g., deliberate destruction of websites). As a community, we therefore need to explore alternative research methodologies that allow us such otherwise impossible research. One promising direction might be the use of alternative testbeds, such as bug bounty sites as a legal proxy to the broader Web [S3]. Another direction could be the development of opt-in mechanisms, similar to a `security.txt`, that allow website operators to explicitly participate in measurement studies.

The Broader Perspective. Stepping back, many of the challenges identified in this thesis cannot be solved solely through technical solutions or written papers. Instead, and we hope this dissertation made it clear, we need to have ongoing discourses and reflections within the research community and exchange with people from the outside.

First, discourse within the community is essential as establishing shared norms and expectations requires open discussions among the affected researchers. Recent efforts focusing on meta-research, such as workshops [8] and broad discussions on open science, reproducibility, and research ethics, are important steps in this direction.

Second, discussions should also take place within individual research domains (like web measurements). Workshops and research visits provide valuable opportunities to reflect on methodological and ethical challenges in one's field. Such exchanges can help researchers and newcomers understand the ethical expectations of their sub-community.

Third, discourse should also take place at the level of research groups. Informal discussions within the labs can play a crucial role in shaping responsible research practices. Such discussions often lead to new ideas and research directions (like [S2]), including methodological innovations and reflecting on research practices (see Chapter 6).

Finally, ethical discourse must also include those who are affected by research. This includes in the Web domain the website operators whose systems may be targeted by measurement studies, but also policymakers and regulators who shape the legal frameworks in which research takes place. Engaging with stakeholders outside the academic community, as well as interdisciplinary collaboration with experts from domains such as law and social sciences, can help ensure that security and privacy research remains both scientifically valuable and socially responsible.

8.3 Concluding Thoughts

As the Web continues to evolve, the responsibility and need for reproducibility of researchers' experiments will only grow. Building a culture of ethical reflection, and transparency will therefore be essential for the future of web security and privacy research. Returning to the motivation of this thesis, confidence in science, we deeply believe that responsible and reproducible research will help keep and increase this confidence in science. At the same time we know, there are more potential challenges to this confidence, not only by issues in cybersecurity research but also by broader developments such as collusion rings, excessive or uncritical use of AI in research, and plagiarism. Addressing these challenges requires collective effort across scientific disciplines.

> *"Whoever is careless with the truth in small matters cannot be trusted with important matters."* — Albert Einstein

As researchers, we constantly make small decisions: we choose our methods, design experiments, and decide which aspects of our work we put in the limited pages we have for publications. Each of these choices shapes how knowledge is produced and communicated. Individually, these decisions seem minor, but collectively they influence the integrity and credibility of research. Reflecting on these choices and keeping truthfulness at the center of our work is therefore essential. Only by doing so can we ensure that our research contributes to scientific progress and strengthens trust in science.

Part IV

Appendix

A AI Usage Declaration

This dissertation was produced with the assistance of AI-based tools. The following statement provides transparency regarding the extent to which these tools were used.

Generative AI tools were used for proofreading, and to enhance style and grammar of existing text. It was not used to generate raw text. Only after section drafts were completed, ChatGPT and Grammarly were used to assist with proofreading and suggesting alternative formulations. No suggested AI-generated formulations were directly adopted. Instead, suggestions were used solely as inspiration for restructuring sentences, with all final wording written by the author himself.

ChatGPT and ScholarLabs were also used as additional engine to source related work supporting traditional methods like keyword search and skimming cited work. No source was included in this dissertation without manual verification of its existence, content, and relevance.

Lastly, ChatGPT was used to generate the images in the beginning of each chapter.

B

A New Web Measurement Tool

B.1 De-Obfuscated Google Analytics Code

In Chapter 4, we talk about Google Analytics. Google Analytics uses programmatically added event listeners for an HTML node to react to events while still allowing CSP to block inline scripts (see Listing 6).

Listing 6 A deobfuscated example of Google Analytics code programmatically setting event listeners.

```
1 ele = document.createElement("script")
2 ele.type = "text/javascript"
3 ele.src = ff.createScriptURL(url)
4 ele.onload = loadHandler
5 ele.onerror = errorHandler
6 ele.setAttribute("nonce", nonce)
7 scr = document.getElementsByTagName("script")[0]
8 scr.parentNode.insertBefore(ele, scr)
```

C

Legal and Ethical Challenges

C.1 3S Scenarios

Here, we list the scenarios presented to the interviewees in Chapter 5. Each participant had the option to ask follow-up questions clarifying technical misunderstandings. For the survey, we slightly modified the wordings to increase clarity, based on insights from the interviews. These versions are included in the full questionnaire [92].

Alice – SQL Injection: Alice checks web servers for vulnerable database queries (e. g., via SQL injection). She uses a function to delay the database response (e. g., the MySQL function "SLEEP"). This allows her to verify whether the server is vulnerable or not.

Bob – Invalid HTTP Request: Bob sends a non-standard HTTP request to a web server. This causes the server to crash *unintentionally*. The result is that the server must now be restarted by the website operator's IT department.

Charlie – Insecure Direct Object Reference: Charlie changes his own user ID in a (1) GET or (2) POST request and to (1) receive and (2) change data from another user.

Daisy – Stored XSS: Daisy exploits a stored XSS (cross-site scripting) vulnerability to deliver its crafted code to potentially *all* users of a website. This code is executed on those users' end devices. It sends a confirmation message back to Daisy's server.

Eve – Path Traversal: Eve modifies a link to a web page to read information that is supposed to be confidential but can be publicly viewed due to server-side configuration issues (e. g., a path traversal).

C.2 Survey Participants

Table C.1 provides statistics on the surveyed operators' demographics and background for Chapter 5. [M] indicates multiple-choice questions for which response counts can sum up to more than 100 %. Percentage values are relative to the total number of survey responses ($n = 119$). For the type of security training received and the number of servers operated (indented lists), percentage values are relative to the number of participants who indicated to have received security training / to operate servers.

Table C.1: Survey participants' demographics and background

	Demographics	n	%
Age	≤ 20	1	0.8
	21–30	11	9.2
	31–40	33	27.7
	41–50	38	31.9
	51–60	21	17.6
	≥ 61	5	4.2
	N/A	10	8.5
Gender[M]	Woman	3	2.5
	Man	101	84.9
	Non-binary	0	0.0
	Self-described	3	2.5
	Prefer not to disclose	10	8.4
	N/A	5	4.2
Security Training[M]	Yes	110	92.4
	University / school	58	48.7
	Employer training	42	35.3
	Certifications	35	29.4
	Other courses	35	29.4
	"Learning by doing"	92	77.3
	Professional network	49	41.2
	Personal network	50	42.0
	Self-taught	106	89.1
	Other	9	7.6
	N/A	10	8.4
	No	8	6.7
	Don't know	1	0.8

	Background	n	%
Responsibility for servers	Yes	118	99.2
	1	3	2.5
	2–5	28	23.5
	6–10	25	21.0
	11–50	40	33.6
	51–100	6	5.0
	101–1,000	9	7.6
	≥ 1,001	5	4.2
	N/A	3	2.5
	No	0	0.0
	Don't know	1	0.8
Main role in server operation	Security Engineer	9	7.6
	System Engineer	23	19.3
	Database Engineer	0	0.0
	Frontend Developer	1	0.8
	Backend Developer	6	5.0
	Full-stack Developer	25	21.0
	DevOps	19	16.0
	Management	14	11.8
	IT Consultant	6	5.0
	Penetration Tester	1	0.8
	Content Creator	2	1.7
	Other	13	10.9
Size of largest company	1	10	8.4
	2–5	6	5.0
	6–10	6	5.0
	11-50	22	18.5
	51–100	5	4.2
	101-1,000	24	20.2
	1,000–10,000	14	11.8
	≥ 10,000	11	9.2
	Don't know	21	17.6
Country of largest company	Germany	36	30.3
	USA	21	17.6
	Austria	5	4.2
	United Kingdom	5	4.2
	Belgium	4	3.4
	Canada	4	3.4
	Switzerland	4	3.4
	Other	36	30.3
	N/A	4	3.4

C.3 Interview Guides

Members of Research Ethics Committees

Table C.2: Interview guide for ethics committee members (part 1)

Part	Questions and Explanations
	Do you consent to participating in this interview, its recording of it, and it being transcribed via Amberscript?
REC	What are currently the main ethical challenges for cyber-security research? What were most difficult cases you had to decide while acting in your capacity as REC member? How many members are there in the ethics committees of the main conferences? How do you recruit members for ethics committees? Do committees remain quite stable in terms of personnel over longer periods of time or is there a lot of fluctuation?
Ethics in Research	Have you ever assessed security studies (planned projects or submitted papers) that used 3S techniques? What were key ethical issues in these studies? From the ethical point of view, how do you assess 3S as a method of data-gathering in security studies? Are there any controversial 3S topics among the members of the committee? How far can security researchers go when testing the behavior of internet infrastructure and websites? Can there be *gray areas* when it comes to studies using 3S techniques and testing of Internet infrastructure? If so, what are these *gray areas*? How should researchers who operate in such *grey areas* be assessed from an ethical perspective? Can you identify absolute *red lines* when it comes to the use of 3S techniques for security research purposes? If so, what are these *red lines*? How would you justify each of these *red lines*? How do you see the relationship between ethical norms regulating scientific research and legal norms? Should security researchers be allowed to do everything that is legal? If not, how would you justify ethical restrictions that go beyond legal restrictions? Should security researchers be allowed to act illegally if it benefits science and/or the broader society? If yes, what would justify such transgressions of legal norms?

Table C.3: Interview guide for ethics committee members (part 2)

Part	Questions and Explanations
Ethics in Research	What should be the absolute normative boundary for such law-transgressive security research? Who should be authorized to decide if a legal transgression by a scientific research team is still acceptable? What kind of institutional safeguards are already in place or should be additionally created? Do you consent to participating in this interview, its recording of it, and it being transcribed via Amberscript?
Improvements	Do you have specific ideas how the use of 3S techniques by security researchers should be regulated to maximize the benefit for all stakeholders (potentially) involved? What would you propose as key stipulations, if asked to design an ethical framework for 3S research? What would you add to the existing ethical framework regulating security research to make 3S studies beneficial? *Explaining our pre-registration proposal...* What do you think about it? How would you re-organize existing institutions tasked with control over security researchers' adherence to norms of scientific ethics? Are there serious flaws in currently existing institutions tasked with control over security researchers' ethics?
Outro	Is there something else you would like to share with us on the topic of ethical regulation of security research - in general or with regard to the use of 3S techniques?

Legal Experts

Table C.4: Interview guide for legal experts. Some questions were only for academics (A), prosecutors (P), or lawyers (L) (part 1)

Part	Questions and Explanations
	Do you consent to participating in this interview, its recording, and it being transcribed via Amberscript?
Intro	Where do you currently see as the most important or biggest challenges in the legal regulation of web security ? (A) What is currently being discussed in this field of legal research? (P) Which security phenomena relevant to criminal law are currently most present in your work as a prosecutor? Are attempts to exploit vulnerabilities of websites currently an important issue in legal debates in your domain? If yes, what exactly is being discussed? Where do you obtain the technical expertise when dealing with questions of legal regulation of digital topics? Are you familiar with the term *server-side scanning*? If yes, how would you define this term?
Criminal Law	What legal risks do you see in server-side scans that are carried out by researchers? What criminal consequences could arise for someone conducting such scans? Have there been cases where criminal prosecution has occurred due to such scientifically motivated tests? (P) (L) Were you involved in any of such cases? What were the results of these proceedings? What forensic means are used in such cases? What defense strategies are chosen by lawyers for those affected by the prosecution? How successful do you estimate these strategies to be? How does prosecution in such cases practically come about? (P) How can a prosecutor get information about an attack on a server? (P) Do authorities focus on filed reports or are there also active investigative against web attacks? (P) What other agencies besides your prosecutor's office are involved in prosecuting such cases?

Table C.5: Interview guide for legal experts. Some questions were only for academics (A), prosecutors (P), or lawyers (L) (part 2)

Part	Questions and Explanations
Criminal Law	Which authorities are responsible for prosecution in such cases? Is there a corresponding specialization in the public prosecutor's offices? Is there a corresponding specialization in police stations or within state or federal police authorities? Who investigates in cross-border cases? Are there lawyers who specialize in such cases or became known for defenses in such cases?
Civil Law	What civil law consequences could arise for someone conducting such tests? Is this relevant for your work as a lawyer (L) / prosecutor (P)? (A) Have there been specific cases where civil lawsuits have been filed against someone conducting such tests? (A) How did these lawsuits come about? (A) What were the results of these proceedings? (A) What forensic means are used in such cases? (A) What strategies were chosen by the lawyers of the defendants? (A) Are there law firms that specialize in such lawsuits?
Legal Misc	Is privacy law relevant for someone conducting such tests? Are there other areas of law relevant for someone conducting such tests? Are there gray areas when conducting such tests? In relation to IT security, are there cases where the benefits for the general public outweigh a potential crime? Looking at the USA, we see they adjusted their *Computer Fraud and Abuse Act*. The attacker's intention is now also considered. The German coalition agreement mentions something similar. Do we currently see a shift in legislation? What is the current status here? How can such legal changes affect legal practice?
International	How does the law deal with cases of cross-border attacks on websites? How important is international law in cases of cross-border attacks on websites? What differences in legal protection of websites are there between different countries? What importance does European law have in this area? What importance does US law have in this area? The USA is not only the cradle of Internet technology, but to this day also the most important place for technological innovation. How does this fact affect the legal regulation of activities on the Internet? How does the fact that many components of the entire Internet infrastructure are not located in Germany or the EU affect the possibilities for legal regulation of activities (especially research activities) on the Internet? Are there different legal approaches in different countries when it comes to cases of web attacks?
Improvements	Do you see a way for researchers to legally conduct 3S? If someone asked you, how would you create a possibility to enable 3S for researchers? *Explain our pre-registration proposal...* What do you think about it?
Outro	In the case of publishing vulnerability analyses, what should researchers pay attention to? Is there something else you would like to share with us on the topic of 3S?

Operators

Table C.6: Interview guide for web operators (part 1)

Part	Questions and Explanations
	Do you consent to participating in this interview, its recording, and it being transcribed via Amberscript?
Background	Please describe in what way you are responsible for a website. What is your daily (primary) job? How is the website maintenance related to your primary job (if related at all)? How did you build the website(s)? How do you maintain the website on a daily (ongoing) basis? How come that you know how to operate a website? Can you tell us a bit about your background? Did you learn programming at school / university? What kind of courses did you attend? Did you learn programming in specialized trainings / courses? What kind of specialized trainings did you attend? Did you teach programming yourself? How did you do that more specifically? What kind of knowledge sources were you using in this process?
IT Security	What are currently the main challenges for operators like you when it comes to hacking? Is the topic of hacking discussed among operators like you and your colleagues? How would you assess your security awareness as an operator? Where do you get the IT security knowledge from? Where do other operators get their IT security knowledge from? What role do interpersonal contacts with other operators play in your acquisition of new knowledge and skills? What is the role of online forums in acquiring IT-related knowledge? Were you ever involved in a vulnerability disclosing process? Can you go into more detail what happened there?
Server-Side Scanning	Are you familiar with the term *server-side scanning* (3S)? If so, how would you define it? Did you ever recognize someone conducting 3S on your server? If yes: Please tell me about your reaction / what you did after you noticed. What is your general opinion about 3S? Should it be allowed for research purposes? Can you describe how you think 3S research is conducted? Should it be allowed for some other purpose?

Table C.7: Interview guide for web operators (part 2)

Part	Questions and Explanations
Server-Side Scanning	Are you familiar with the term *server-side scanning* (3S)? Do you see any problems for researchers doing 3S? Would you sue a researcher who conducted 3S on your IT infrastructure? Would the location of the IP address of the 3S actor make a difference? How far can security research go (in your opinion)? What are *gray areas*? Can you define absolute *red lines*? How would you justify these *red lines*? Do you see problems for you / your company resulting from 3S? What would happen if something breaks (after 3S)? What would happen if data is leaked (after 3S)? Would you analyze or look closer at traces of a 3S attempt? Would you improve your security after noticing 3S?
Improv.	Do you have any specific ideas how we could improve the current situation in a way that 3S is possible for researchers? If you were asked to design a guideline for security researchers, what would you propose? *Explain our pre-registration proposal...* What do you think about it?
Outro	Is there something else you would like to share with us on the topic of 3S?

C.4 3S Assessments in the Survey

Table C.8 provides the results from the survey in Chapter 5 regarding the proposed 3S scenarios.

Table C.8: Surveyed operators' (n = 119) general assessment of 3S, all scenarios and more or less invasive variants, and our pre-registration proposal

		Likert scores									
Q#	Scenario	Comfortable (5)	Somewhat comfortable (4)	Neither comf. nor uncomf. (3)	Somewhat uncomfortable (2)	Uncomfortable (1)	Not shown	N/A	mean	std	median
		%	%	%	%	%	%	%			n
Q6	3S General	36.1	21.8	9.2	13.4	19.3	–	0	3.42	1.55	4.0
Q8	Pre-registration	37.8	35.3	14.3	3.4	9.2	–	0	3.89	1.22	4.0
A1	Alice (base)	51.3	18.5	7.6	8.4	14.3	–	0	3.84	1.48	5.0
A2	Alice (more inv.)	24.4	13.4	12.6	13.4	13.4	22.7	0	2.54	1.91	3.0
A3	Alice (less inv.)	1.7	2.5	7.6	6.7	11.8	69.7	0	0.66	1.21	0
B1	Bob (base)	34.5	20.2	7.6	11.8	26.1	–	0	3.25	1.64	4.0
B2	Bob (more inv.)	7.6	14.3	8.4	16.0	16.0	37.8	0	1.68	1.71	1.0
B3	Bob (less inv.)	1.7	8.4	10.1	7.6	17.6	54.6	0	1.05	1.44	0
C1	Charlie (base)	50.4	16.0	10.9	5.9	16.8	–	0	3.77	1.53	5.0
C2	Charlie (more inv.)	12.6	15.1	6.7	11.8	31.1	22.7	0	1.98	1.75	1.0
C3	Charlie (less inv.)	9.2	3.4	6.7	1.7	12.6	66.4	0	0.96	1.66	0
D1	Daisy (base)	23.5	16.0	10.9	16.0	33.6	–	0	2.80	1.61	3.0
D2	Daisy (more inv.)	10.1	9.2	4.2	9.2	17.6	49.6	0	1.36	1.76	1.0
D3	Daisy (less inv.)	10.9	10.1	5.9	15.1	18.5	39.5	0	1.61	1.76	1.0
E1	Eve (base)	47.1	20.2	10.1	7.6	14.3	–	0.8	3.76	1.50	4.0
E2	Eve (more inv.)	26.1	15.1	8.4	12.6	15.1	22.7	0	2.56	1.95	2.0
E3	Eve (less inv.)	2.5	0.8	9.2	6.7	12.6	68.1	0	0.70	1.23	0

D

Ethical Practices at Conference

D.1 Survey Tool

As one part of the interview in Chapter 7, we let the interviewees share their screen and go through a survey tool. The tool shows several suggested interventions in form of items that the participant drags and drops in buckets labeled with a Likert scale: Strongly agree – Somewhat agree – Neither agree nor disagree – Somewhat disagree – Strongly disagree. The introduction text and items of the tool are listed below.

> Would you agree that the following ideas for ethics at conferences should be implemented (ignore feasibility)? Please place the items shown below by drag & drop into response boxes depending on your assessment of the given item. New ideas appear after one is placed.

- Publishing meta-reviews summarizing reviewers' discussions on ethical considerations.
- Consolidating research ethics committees from multiple conferences into a single board.
- Providing ethical training for reviewers.
- Introducing a Profound Ethical Considerations badge, similar to artifact evaluations.
- Making an ethics section mandatory for all research papers.
- Including external ethics experts in ethics boards at the conferences to ensure perspectives beyond our research community.
- Providing educational material on ethics in form of categorized topics (e.g., human subjects, responsible disclosure, etc.).
- Mentoring system for research ethics committee

Would you agree that the following instruments should be incorporated into the review forms for our conference? Please place the items shown below by drag & drop into response boxes depending on your assessment of the given item.
Ethics Score: The "score" that the author receives in the review
Internal Ethics Concerns: Internal documentation about the concerns for the REC and PC
Feedback for Author: The feedback that the authors get with the review

- Ethics Score: Binary (Ethical / Concerns)
- Ethics Score: Numeric
- Ethics Score: Severity (e.g., none, minor, moderate, severe)
- Internal Ethics Concerns / Considerations: Selectable Pre-formulated Categories
- Internal Ethics Concerns / Considerations: Free Text
- Feedback for Author: Pre-formulated Categories for Required Author's Action
- Feedback for Author: Free Text
- Reviewer Self-Assessment (Are you confident in evaluating the ethics of this paper)
- Ethics fields in review must be mandatory

D.2 Keyword Analysis

To understand how the topic of research ethics has evolved over time within the S&P research community, we conducted a keyword analysis of security and privacy papers published at top-tier conferences during the period 2000-2024. In contrast to previous work by Zhang et al. [268], statistical features such as TF-IDF weighting and word co-occurrences were used to ensure the robustness of the extracted keyword list. Our results indicate an overall increase in awareness of ethical issues, particularly within the botnet sub-community.

Method

Data collection

We leveraged DBLP and collected papers from the top-tier security and privacy conferences, namely CCS, IEEE S&P, NDSS and USENIX Security, from the period 2000 to 2024. The crawler that we used for data collection worked very reliably for all venues except the older S&P conferences, where detection rates often fell below 30%. To ensure the validity of our measurement, we began the analysis in the year 2000. With this restriction, fewer than 100 papers are missing from a total data set of 8489 papers. The collected papers were parsed into raw text using docling [13].

Data analysis

We created several keyword lists on different topics, *General Ethics*, *Human Subjects*, and *Botnet*, with each list covering the relevant topic as comprehensively as possible.

General Ethics includes general references to ethical aspects, whereas *Human Subjects* covers references to working with human participants (e.g., users or developers).

First, the papers were pre-filtered based on their chapter headings. For the *General Ethics* dimension, for example, all chapters containing the word *ethic* in their heading were extracted. These texts were then used to create the keyword list for the corresponding topic. Statistical features such as TF-IDF weighting and word co-occurrences were applied to extract the final keyword list. Both methods ensure the representativity, the specificity, as well as the exhaustivity of the lists and are widely used in the field of keyword extraction [74]. This co-occurrences-based approach allows a more robust keyword analysis and reduces false positives compared to prior work (e.g., [268]), as it relies on groups of related terms rather than individual keywords. The topics examined and their associated keyword lists are shown in Table D.1. In order to indicate the presence of each topic over time, all papers containing at least one word from the corresponding keyword list were counted for each topic and each year.

Results

The results are displayed in the figures below (Figure D.1 - D.3). In addition to the data itself, the graphs also show historically interesting points that may have influenced the treatment of ethical issues in S&P papers. These include the introduction of the Menlo Report, the first mention of ethics-related content in the CfP, and what we have termed the *botnet controversy* described above.

Figure D.1 visualizes the data of the *General Ethics* dimension. Across all four conferences, a noticeable, albeit non-monotonous, increase can be seen within the studied period. Since 2012, at least one paper with ethical content has been published at each of the conferences examined. However, there is no clear correlation between the trend and the historical points marked. There are both upward and downward trends during the botnet controversy. With regard to the Menlo Report and the first mention of ethical content in the CfP, no clear trends can be identified in the following years.

The data of the *Human Subjects* topic is displayed in Figure D.2. Similar to the *General Ethics* dimension, a higher presence of the topic can also be seen here in recent years. However, *General Ethics* and *Human Subjects* did not develop in the same way. There were noticeable sharp declines at USENIX 2014 and S&P 2013. No overarching patterns can be identified with regard to the historical markers either.

In addition to these general ethical categories, we also took a closer look at papers dealing with the topic of botnets (Figure D.3). With the exception of a slump in 2011 at the NDSS, the number of botnet papers across all conferences rose steadily during the botnet controversy. Around 2012, interest in the topic began to decline sharply, with the proportion of papers containing botnet-related content fluctuating around ten percent in recent years. The NDSS is once again an exception here. Here, the presence of the topic is at a slightly higher level, between ten and 20 percent. However, while the overall presence of the topic has declined since the years following the botnet controversy, an almost opposite dynamic can be observed in botnet papers that deal with ethics-related content. With the exception of the CCS, ethics played hardly any role in the botnet field at the end of the 2000s, but in the years that followed, there

were several years in which ethics were widely discussed within this sub-community, such as at CCS 2024, USENIX 2023, and NDSS 2019. This points to an increased ethical awareness in this sub-community.

Limitations

One of the biggest problems with keyword analysis, as carried out here, is that the results depend on the keyword lists used. Although TF-IDF and word co-occurrences can ensure a certain quality in terms of representativeness and specificity of the keywords, the use of these methods does not provide a clear result. For example, one particularly representative word pair for texts whose headline contains *botnet* is *IP addresses*. However, this word pair can be found in all kinds of areas of computer science. The approach used here cannot compare different thematic areas of computer science in order to further increase the specificity of the keywords and identify clearly distinct terms.

It should also be emphasized that no causal relationships can be inferred from the keyword analysis. On the one hand, this is due to the method itself and, on the other hand, to the way in which the community might react to the emerging treatment of ethical issues. It is conceivable, for example, that there will be a delay in responding to novel interventions such as the Menlo Report and that conferences will influence each other in their approach to the topic of ethics. This makes it extremely difficult to trace the results back to specific historical events.

Furthermore, it should be noted that the method used here is not suitable for measuring arbitrary topics. Attempts were also made to create keyword lists for the areas of *attack* and *measurement* research. However, the chapters titled with the respective terms are far too numerous and cover too many subtopics to extract a specific keyword list. In order to use this method, a more precise understanding of possible subcommunities and sub-topics must therefore be available in advance. Such sub-topics, whose development could explain the increase in papers with an ethical reference, cannot be found using this method.

	Heading Keyword	Keyword List
General Ethics	ethic	review board, institutional review, board irb, irb approval, board erb, ethical considerations, ethics committee, ethical concerns, ethical issues, institution irb, ethical implications, ethical principles, ethical frameworks, menlo report, belmont report, public interest
Human Subject	ethic	consent form, explicit consent, human subjects, real users, involve human
Botnet	botnet	botnet, infected machines, infected hosts

Table D.1: Keyword lists for the topics examined

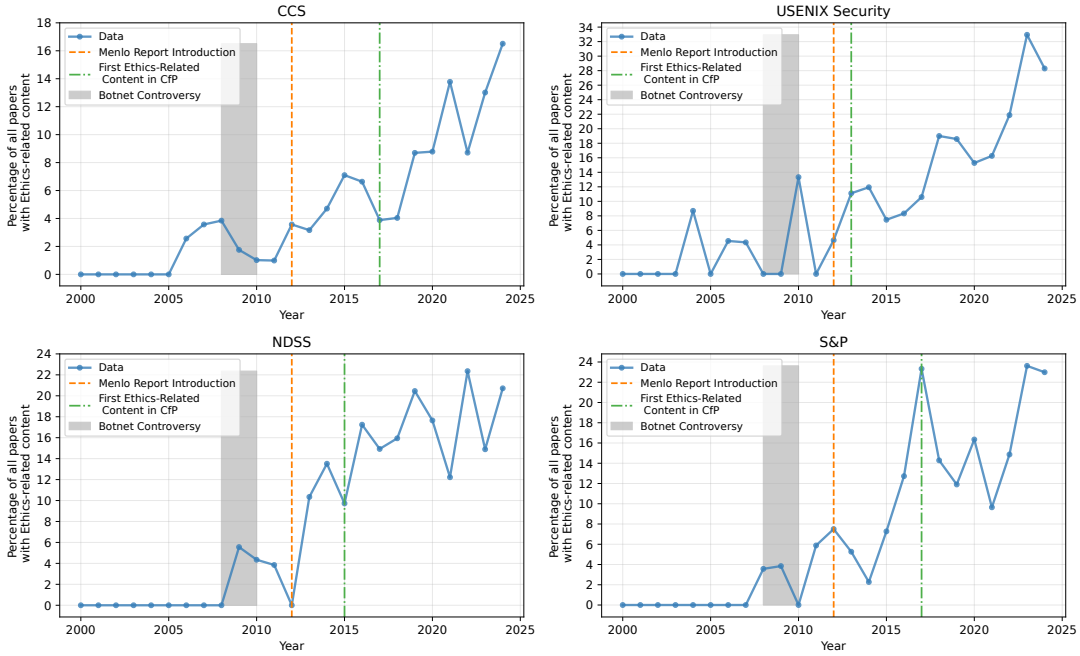


Figure D.1: Paper with ethics-related content over time, broken down by conference

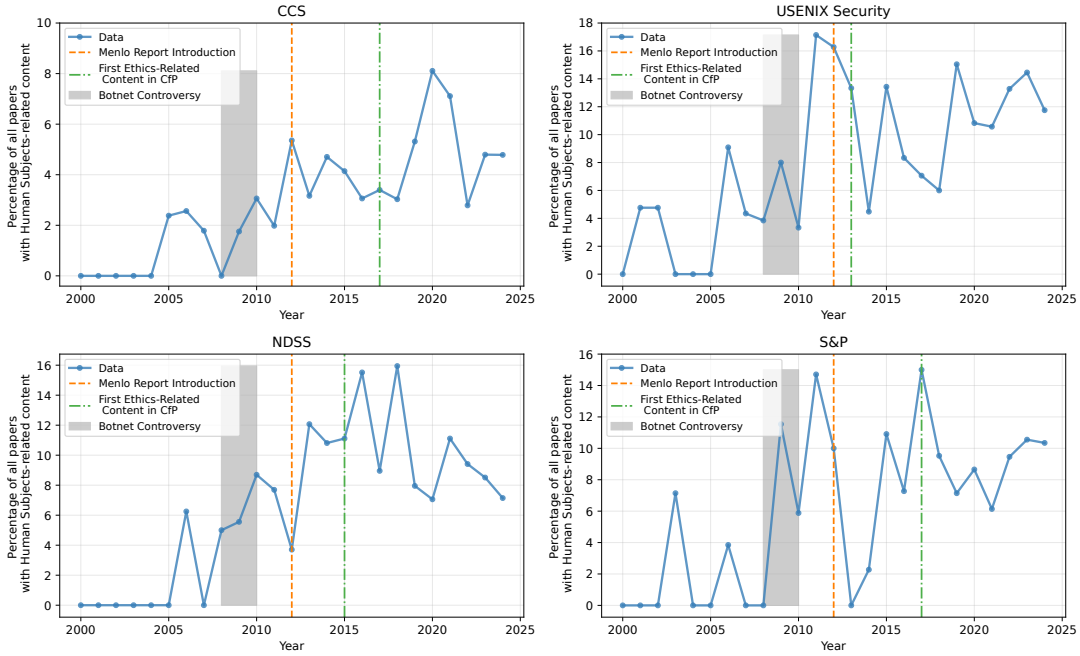


Figure D.2: Paper with human subjects-related content over time, broken down by conference

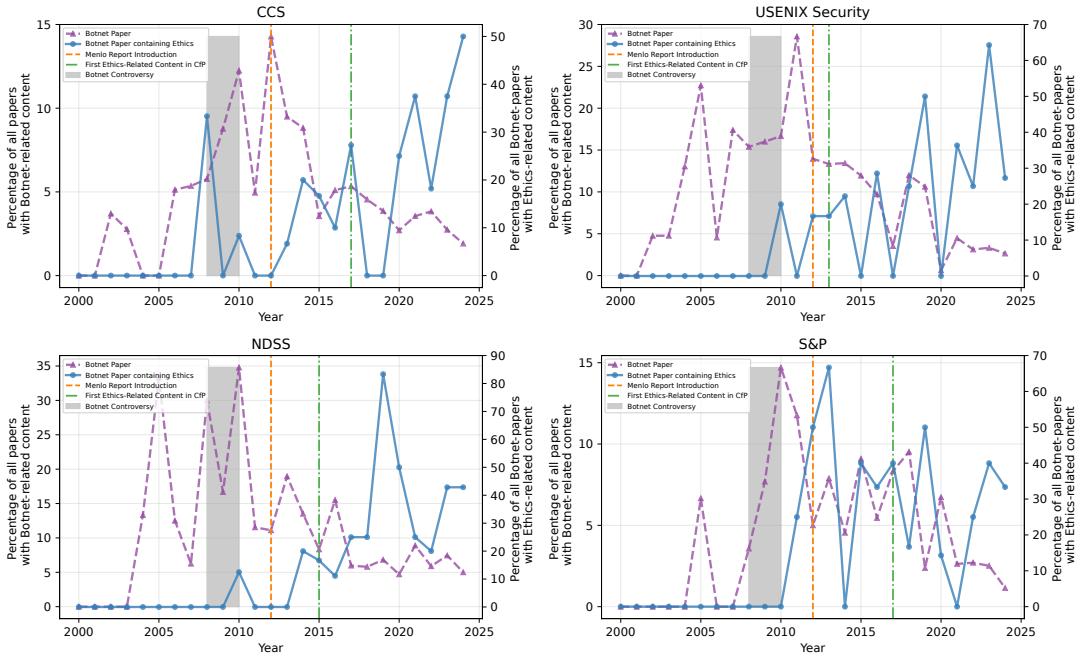


Figure D.3: Paper with botnet-related content (dashed line), as well as the share of all botnet-related papers dealing with ethical topics (solid line), over time, broken down by conference

D.3 Interview Guide

Introduction	Before we dive in, how would you describe your overall experience with ethics-procedures at academic conferences as a PC chair or member and as an author?
Ethics Procedures	Let's start right into the interview: You are/were the chair of CONFERENCE. CONFERENCE has/had a research ethics committee supporting the review phase. Could you describe this procedure for a submitted paper?
Ethics Interventions	We explain our definition of research ethics intervention: Everything at a conference that influences how we consider ethics in our research, e.g., ethics reviews, guidelines, other information provided. We will go through some ethics interventions that conferences had implemented in the past. Please tell me your opinion about each, is it helpful or not? • Research Ethics Committees: S&P and USENIX have a REC since 2022, NDSS followed 2025, what do you think about them? • Mandatory Ethics Section: USENIX made ethics sections mandatory for all paper in 2025. What do you think about this? • IRB Decision Disclosure: All conferences ask for information about IRB decision for papers. What do you think about this? • Responsible Disclosure Requirements: S&P and USENIX explicitly recommend how to conduct responsible disclosure. What do you think about this? • Ethics Categories: Some conferences list what kind of research and methods require ethical considerations. For example, human subject research, research on live practices, or well-being of the research team. What do you think about this (ethical categorization in general)? • Right to reject: Some conferences mention explicitly that they might reject unethical paper. What do you think about this? • Ethics Frameworks: Conferences suggested various frameworks to evaluate ethics in the past. For example, evaluating harm vs. benefit or a stakeholder analysis. Are such frameworks helpful or should authors be free in deciding how they discuss ethics in their papers? Do you know of any other interventions implemented to ensure ethical considerations during paper writing and reviewing at the conference? Please explain the details of the intervention to me.
Survey Tool	In this part of the interview, we let the interviewees share their screen and go through a survey tool (see Appendix D.1).

Table D.2: Interview guideline (part 1)

Reasoning	What was the reasoning behind the decision to implement ethics practices and interventions? Were there discussions or debates in the PC about this implementation and alternatives? If so, what were the main arguments? Have you or your predecessors tried out other instruments or frameworks? If so, which ones, and how did they perform? Did the PC consider insights from other fields (e.g., sociology, philosophy, medical research, psychology) to design these interventions?
Goal	In your view, what are the primary goals of implementing research ethics practices at conferences? Are these goals already reached or is it an ongoing process towards these goals? (Optional) How well do you think the current approach achieves these goals? What could be improved with the current practice? Are there any aspects of the current ethics practices at conferences that you don't like? What are your thoughts on the development of ethics considerations in our community? (Optional) Do we need IRB approvals and committees at conferences at the same time?
Importance of Ethics	What are your thoughts on how to handle unethical papers that have a great contribution? Stepping away from your role as PC chair, how do you personally see the importance of ethics in research compared to other aspects, such as novelty or technical correctness?
Ending	Is there anything else you'd like to add about the ethics review procedure?

Table D.3: Interview guideline (part 2)

Bibliography

Author's Papers for this Thesis

[P1] Hantke, F., Calzavara, S., Wilhelm, M., Rabitti, A., and Stock, B. You Call This Archaeology? Evaluating Web Archives for Reproducible Web Security Measurements. In: *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 2023. DOI: 10.1145/3576915.3616688.

[P2] Hantke, F., Roth, S., Mrowczynski, R., Utz, C., and Stock, B. Where Are the Red Lines? Towards Ethical Server-Side Scans in Security and Privacy Research. In: *2024 IEEE Symposium on Security and Privacy (SP)*. 2024. DOI: 10.1109/SP54263.2024.00104.

[P3] Hantke, F., Snyder, P., Haddadi, H., and Stock, B. Web Execution Bundles: Reproducible, Accurate, and Archivable Web Measurements. In: *34th USENIX Security Symposium (USENIX Security 25)*. 2025.

[P4] Hantke, F., Mrowczynski, R., and Stock, B. Reflection, Education, Consistency: Towards Best Ethics Practices At Security And Privacy Conferences. In: *Proceedings of the 2026 ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 2026.

Other Papers of the Author

[S1] El Hajj Chehade, S., Hantke, F., and Stock, B. 403 Forbidden? Ethically Evaluating Broken Access Control in the Wild. In: *2025 IEEE Symposium on Security and Privacy (SP)*. 2025. DOI: 10.1109/SP61157.2025.00252.

[S2] Mustafa, A., Rautenstrauch, J., Hantke, F., Agarwal, S., Calzavara, S., and Stock, B. LeakyLinks: Measuring the Security and Privacy Risks of URL Scanning Services. In: *2026 IEEE Symposium on Security and Privacy (SP)*. 2026.

[S3] Decker, P. and Hantke, F. VDPCollect: Vulnerability Disclosure Programs as a Complement to Web Security Measurements. In: *Proceedings of the 21st ACM Asia Conference on Computer and Communications Security (AsiaCCS)*. 2026.

[S4] Hantke, F. and Stock, B. HTML Violations and Where to Find Them: A Longitudinal Analysis of Specification Violations in HTML. In: *Proceedings of the 22nd ACM Internet Measurement Conference (IMC)*. 2022. DOI: 10.1145/3517745.3561437.

Other references

[1] Abel, R. L., Hammerslev, O., Sommerlad, H., and Schultz, U., eds. *Lawyers in 21st-Century Societies: Volume 1: National Reports*. Hart, 2020.

[2] Abel, R. L., Hammerslev, O., Sommerlad, H., and Schultz, U., eds. *Lawyers in 21st-Century Societies: Volume 2: Comparisons and Theories*. Hart, 2022.

[3] ACM. *Artifact Review and Badging – Current*. 2020. URL: https://www.acm.org/publications/policies/artifact-review-and-badging-current (visited on 01/27/2026).

[4] Ahmad, S. S., Dar, M. D., Zaffar, M. F., Vallina-Rodriguez, N., and Nithyanand, R. Apophanies or Epiphanies? How Crawlers Impact Our Understanding of the Web. In: *Proceedings of The Web Conference 2020*. 2020. DOI: 10.1145/3366423.3380113.

[5] Ainsworth, S. G., Nelson, M. L., and Sompel, H. V. de. Only one out of five archived web pages existed as presented. In: *ACM HT*. 2015. DOI: 10.1145/2700171.2791044.

[6] Alexander, C. S. and Becker, H. J. The use of vignettes in survey research. *Public Opinion Quarterly* 42 (1978).

[7] Alsum, A., Weigle, M. C., Nelson, M. L., and Sompel, H. V. de. Profiling web archive coverage for top-level domain and content language. In: *International Conference on Theory and Practice of Digital Libraries*. 2013. DOI: 10.1007/978-3-642-40501-3_7.

[8] Ananta Soneji and Moritz Schloegel. *MetaCRiSP 2026*. URL: https://metacrisp.org/ (visited on 02/05/2026).

[9] Apel, K.-O. *Diskurs und Verantwortung: Das Problem des Übergangs zur postkonventionellen Moral*. Suhrkamp, 1990.

[10] ATLAS.ti Scientific Software Development GmbH. *ATLAS.ti*. 2026. URL: https://atlasti.com (visited on 03/09/2026).

[11] Atzmüller, C. and Steiner, P. M. Experimental Vignette Studies in Survey Research. *Methodology* 6, 3 (2010). DOI: 10.1027/1614-2241/a000014.

[12] Auchard, E. *Researcher Found Security Holes at Smartphone-Only Bank N26*. 2022. URL: https://web.archive.org/web/20220929184415/https://www.reuters.com/article/us-cyber-fintech-n26-idUSKBN14H1EM (visited on 03/09/2026).

[13] Auer, C. et al. Docling technical report. *arXiv preprint* (2024). DOI: 10.48550/arXiv.2408.09869.

[14] Bailey, M., Dittrich, D., Kenneally, E., and Maughan, D. The Menlo Report. In: *2012 IEEE Symposium on Security and Privacy (SP)*. 2012. DOI: 10.1109/MSP.2012.52.

[15] Balaban, S. et al. *Rechtslage der IT-Sicherheitsforschung*. 2022. URL: https://sec4research.de/assets/Whitepaper.pdf (visited on 03/07/2026).

[16] Bauer, L. and Pellegrino, G. *USENIX Security '25 Ethics Guidelines*. 2024. URL: https://www.usenix.org/conference/usenixsecurity25/ethics-guidelines (visited on 11/11/2024).

[17] Bauer, L. and Pellegrino, G. *Message from the USENIX Security '25 Program Co-Chairs*. 2025. URL: https://www.usenix.org/sites/default/files/sec25_message.pdf (visited on 01/13/2026).

[18] Bauer, L. and Pellegrino, G. *USENIX Security '25 Call for Papers*. URL: https://www.usenix.org/conference/usenixsecurity25/call-for-papers (visited on 09/28/2024).

[19] Bausell, R. B. *The Problem with Science: The Reproducibility Crisis and What to Do About It*. Oxford University Press, 2021.

[20] Begley, C. G. and Ellis, L. M. Raise standards for preclinical cancer research. *Nature* 483, 7391 (2012). DOI: 10.1038/483531a.

[21] Bélanger, F. and Crossler, R. E. Privacy in the Digital Age: a Review of Information Privacy Research in Information Systems. *MIS quarterly* 35, 4 (2011).

[22] Bensalim, S., Klein, D., Barber, T., and Johns, M. Talking about my generation: targeted dom-based xss exploit generation using dynamic data flow analysis. In: *Proceedings of the 14th European Workshop on Systems Security*. 2021. DOI: 10.1145/3447852.3458718.

[23] Bentham, J. *An Introduction to the Principles of Morals and Legislation*. The Athlone Press, 1970.

[24] Berman, P. S. *Global Legal Pluralism: a Jurisprudence of Law beyond Borders*. Cambridge University Press, 2012.

[25] Berners-Lee, T. *The Original HTTP as defined in 1991*. URL: https://www.w3.org/Protocols/HTTP/AsImplemented.html (visited on 02/06/2026).

[26] Bogner, A., Littig, B., and Menz, W., eds. *Interviewing Experts*. Palgrave Macmillan, 2009.

[27] Bohlander, M. *Principles of German Criminal Law*. 1st ed. Hart, 2008.

[28] Bollinger, D., Kubicek, K., Cotrini, C., and Basin, D. Automating cookie consent and GDPR violation detection. In: *31st USENIX Security Symposium (USENIX Security 22)*. 2022.

[29] Bonneau, J. *Oakland REC Annual Summary 2022*. 2022. URL: https://docs.google.com/document/d/15x5Qd1UTaoMSRZgRRvurPbgRg4SWWLQKhG41ouYV0TY (visited on 07/24/2023).

[30] Brandt, A. M. Racism and research: the case of the tuskegee syphilis study. *Hastings center report* (1978). DOI: 10.2307/3561468.

[31] Brave Software Inc. *Brave Browser*. https://brave.com.

[32] Brave Software Inc. *pagegraph-crawl*. URL: https://github.com/brave/pagegraph-crawl.

[33] Brave Software Inc. *pagegraph-query*. URL: https://github.com/brave-experiments/pagegraph-query.

[34] Brunelle, J. F., Kelly, M., SalahEldeen, H., Weigle, M. C., and Nelson, M. L. Not all mementos are created equal: measuring the impact of missing resources. *Int. J. Digit. Libr.* 16, 3-4 (2015). DOI: 10.1007/s00799-015-0150-6.

[35] Bugliesi, M., Calzavara, S., Focardi, R., and Khan, W. Cookiext: patching the browser against session hijacking attacks. *J. Comput. Secur.* 23, 4 (2015). DOI: 10.3233/JCS-150529.

[36] Burnett, S. and Feamster, N. Encore: Lightweight Measurement of Web Censorship with Cross-Origin Requests. In: *SIGCOMM 2015*. ACM, 2015. DOI: 10.1145/2785956.2787485.

[37] Butler, K. and Thomas, K. *Message from the USENIX Security '22 Program Co-Chairs*. 2022. URL: https://www.usenix.org/sites/default/files/sec22_message.pdf (visited on 01/13/2026).

[38] Buyukkayhan, A. S., Gemicioglu, C., Lauinger, T., Oprea, A., Robertson, W., and Kirda, E. What's in an Exploit? An Empirical Analysis of Reflected Server XSS Exploitation Techniques. In: *23rd International Symposium on Research in Attacks, Intrusions and Defenses*. USENIX Association, 2020.

[39] Byers, J. W. *Encore: Lightweight Measurement of Web Censorship with Cross-Origin Requests – Public Review*. 2015. URL: https://conferences.sigcomm.org/sigcomm/2015/pdf/reviews/226pr.pdf (visited on 02/05/2026).

[40] Calzavara, S., Roth, S., Rabitti, A., Backes, M., and Stock, B. A Tale of Two Headers: A Formal Analysis of Inconsistent Click-Jacking Protection on the Web. In: *29th USENIX Security Symposium (USENIX Security 20)*. 2020.

[41] Calzavara, S., Urban, T., Tatang, D., Steffens, M., Stock, B., et al. Reining in the web's inconsistencies with site policy. In: *Proceedings of 2021 Network and Distributed System Security Symposium (NDSS)*. 2021. DOI: 10.14722/ndss.2021.23091.

[42] Camerer, C. F. et al. Evaluating replicability of laboratory experiments in economics. *Science* 351, 6280 (2016). DOI: 10.1126/science.aaf0918.

[43] Centre for Cyber Security Belgium. *Vulnerability Reporting to the CCB*. 2023. URL: https://ccb.belgium.be/en/vulnerability-reporting-ccb (visited on 05/09/2025).

[44] CERN. *A short history of the Web*. URL: https://home.cern/science/computing/birth-web/short-history-web (visited on 02/06/2026).

[45] Chaos Computer Club. *Clause 202c of German penal code endangers German IT industry*. 2008. URL: https://www.ccc.de/en/updates/2008/stellungnahme202c (visited on 05/30/2023).

[46] Chen, Q., Snyder, P., Livshits, B., and Kapravelos, A. Detecting filter list evasion with event-loop-turn granularity javascript signatures. In: *2021 IEEE Symposium on Security and Privacy (SP)*. 2021. DOI: 10.1109/SP40001.2021.00007.

[47] Chrome DevTools. *Protocol*. URL: https://chromedevtools.github.io/devtools-protocol (visited on 06/05/2024).

[48] Cockburn, A., Gutwin, C., and Dix, A. HARK No More: On the Preregistration of CHI Experiments. In: *CHI '18*. ACM, 2018. DOI: 10.1145/3173574.3173715.

[49] Cortier, V. and Lin, Z. *Request for feedback on the ACM CCS 2026 Draft CfP (via GitHub issues)*. 2025. URL: https://github.com/ACM-CCS-2026/CfP (visited on 01/14/2026).

[50] Crandall, J. R., Crete-Nishihata, M., and Knockel, J. Forgive Us Our SYNs: Technical and Ethical Considerations for Measuring Internet Filtering. In: *SIGCOMM Workshop on Ethics in Networked Systems Research*. ACM, 2015. DOI: 10.1145/2793013.2793021.

[51] curl. *curl*. https://curl.se/.

[52] *CVE-2021-36539*. CVE-2021-36539. URL: https://www.cve.org/CVERecord?id=CVE-2021-36539 (visited on 11/05/2024).

[53] *CWE - CWE-639: Authorization Bypass Through User-Controlled Key (4.15)*. URL: https://cwe.mitre.org/data/definitions/639.html (visited on 02/05/2026).

[54] dakitu. *Nord Security Disclosed on HackerOne: IDOR allow access to payments data of any user*. HackerOne. URL: https://hackerone.com/reports/751577 (visited on 02/05/2026).

[55] Degeling, M., Utz, C., Lentzsch, C., Hosseini, H., Schaub, F., and Holz, T. We value your privacy... now take some cookies: measuring the gdpr's impact on web privacy. In: *Proceedings of 2018 Network and Distributed System Security Symposium (NDSS)*. 2018. DOI: 10.14722/ndss.2019.23378.

[56] Demir, N. Better Web Measurements: Enhancing Security and Privacy Research. PhD thesis. Dissertation, Karlsruhe, Karlsruher Institut für Technologie (KIT), 2023, 2023.

[57] Demir, N., GroSSe-Kampmann, M., Urban, T., Wressnegger, C., Holz, T., and Pohlmann, N. Reproducibility and Replicability of Web Measurement Studies. In: *Proceedings of the ACM Web Conference 2022*. 2022. DOI: 10.1145/3485447.3512214.

[58] Dirksen, A., Giessler, S., Erz, H., Johns, M., and Fiebig, T. Don't Patch the Researcher, Patch the Game: A Systematic Approach for Responsible Research via Federated IRBs. In: *Proceedings of the New Security Paradigms Workshop*. 2024. DOI: 10.1145/3703465.3703475.

[59] Disconnect. *Disconnect*. https://disconnect.me/.

[60] Dittrich, D., Bailey, M., and Dietrich, S. Have we crossed the line? The growing ethical debate in modern computer security research. In: *Proceedings of the 2009 on ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 2009.

[61] Dittrich, D., Bailey, M., and Dietrich, S. Towards community standards for ethical behavior in computer security research. *University of Washington, Michael Bailey, University of Michigan, Sven Dietrich, Stevens Institute of Technology, Stevens CS Technical Report 20091* (2009).

[62] Dittrich, D., Bailey, M., and Dietrich, S. Building an active computer security ethics community. In: *2010 IEEE Symposium on Security and Privacy (SP)*. 2010. DOI: 10.1109/MSP.2010.199.

[63] Dittrich, D., Kenneally, E., and Bailey, M. Applying ethical principles to information and communication technology research: A companion to the Menlo Report. *Available at SSRN 2342036* (2013).

[64] Doupé, A., Cavedon, L., Kruegel, C., and Vigna, G. Enemy of the State: A State-Aware Black-Box Web Vulnerability Scanner. In: *21st USENIX Security Symposium (USENIX Security 12)*. 2012.

[65] Durumeric, Z., Wustrow, E., and Halderman, J. A. ZMap: Fast Internet-wide Scanning and Its Security Applications. In: *22nd USENIX Security Symposium (USENIX Security 13)*. Washington, DC, USA, 2013.

[66] EasyList. *Overview*. https://easylist.to/.

[67] Englehardt, S. and Narayanan, A. Online tracking: a 1-million-site measurement and analysis. In: *Proceedings of the 2016 on ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 2016. DOI: 10.1145/2976749.2978313.

[68] Eriksson, B., Pellegrino, G., and Sabelfeld, A. Black Widow: Blackbox Data-Driven Web Scanning. In: *2021 IEEE Symposium on Security and Privacy (SP)*. 2021. DOI: 10.1109/SP40001.2021.00022.

[69] Eshleman, A. Moral responsibility. In: *Stanford Encyclopedia of Philosophy*. 2008.

[70] European Parliament and the Council of the European Union, T. *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. 2016.

[71] Evertz, J. et al. Chasing Shadows: Pitfalls in LLM Security Research. In: *Proceedings of 2025 Network and Distributed System Security Symposium (NDSS)*. 2025. DOI: 10.14722/ndss.2026.241749.

[72] Fernandez-de-Retana, A., Rautenstrauch, J., Santos-Grueiro, I., and Stock, B. A Permissions Odyssey: A Systematic Study of Browser Permissions on Modern Websites. In: 2025. DOI: 10.1145/3730567.3764489.

[73] Finch, J. The Vignette Technique in Survey Research. *Sociology* 21, 1 (1987). DOI: 10.1177/003803858702100 1008.

[74] Firoozeh, N., Nazarenko, A., Alizon, F., and Daille, B. Keyword extraction: issues and methods. *Natural Language Engineering* 26, 3 (2020).

[75] Flick, U. *An Introduction to Qualitative Research*. 7th ed. Sage, 2023.

[76] Gamero-Garrido, A., Savage, S., Levchenko, K., and Snoeren, A. C. Quantifying the Pressure of Legal Risks on Third-Party Vulnerability Research. In: *Proceedings of the 2017 on ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 2017. DOI: 10.1145/3133956.3134047.

[77] Garfinkel, S. L. IRBs and security research: myths, facts and mission creep. In: *Usability, Psychology, and Security 2008 (UPSEC 08)*. USENIX Association, 2008.

[78] German Federal Ministry of Justice. *German Civil Code (Bürgerliches Gesetzbuch – BGB)*. Non-binding English translation: https://www.gesetze-im-internet.de/englisch_bgb/englisch_bgb.html. 2021.

[79] German Federal Ministry of Justice. *German Criminal Code (Strafgesetzbuch – StGB)*. Non-binding English translation: https://www.gesetze-im-internet.de/englisch_stgb/index.html. 2021.

[80] Gilbertson, S. *Germany Outlaws Hacking, Cripples Security Industry*. 2007. URL: https://www.wired.com/2007/08/germany-outlaws-hacking-cripples-security-industry/ (visited on 02/05/2026).

[81] Glaser, B. G. and Strauss, A. L. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Aldine Publishing Company, 1967.

[82] GNU Project. *GNU Wget*. https://www.gnu.org/software/wget/.

[83] Google, Inc. *Puppeteer*. (Visited on 06/05/2024).

[84] gorhill. *uBlock Origin*. https://github.com/gorhill/uBlock.

[85] GroSS, A. *The CDU's leaky campaign app*. 2021. URL: https://www.berliner-zeitung.de/en/the-cdus-leaky-campaign-app-li.176310 (visited on 07/23/2023).

[86] Habermas, J. *Justification and Application: Remarks on Discourse Ethics*. Polity Press, 1993.

[87] Hantke, F. *You Call This Archaeology? – Code Base*. 2023. URL: https://github.com/cispa/internet-archive-study (visited on 03/04/2026).

[88] Hantke, F. *Conference Ethics Supplementary*. 2026. URL: https://github.com/FHantke/Ethical-Practices-at-Conference-Supplementary-Material (visited on 03/12/2026).

[89] Hantke, F. *Security and Privacy Ethics Wiki*. 2026. URL: https://anonymous.4open.science/r/research-ethics-wiki-21C5 (visited on 01/09/2026).

[90] Hantke, F. *.web Datasets*. https://doi.org/10.5281/zenodo.14760512.

[91] Hantke, F. *WebREC*. https://doi.org/10.5281/zenodo.14718651.

[92] Hantke, F., Roth, S., Mrowczynski, R., Utz, C., and Stock, B. *Where Are the Red Lines? Towards Ethical Server-Side Scans in Security and Privacy Research – Supplementary Material*. 2023. URL: https://github.com/cispa/Ethical-Server-Side-Scanning (visited on 08/03/2023).

[93] Harán, J. M. *A Vulnerability in Instagram Exposes Personal Information of Users*. WeLiveSecurity ESET. 2019. URL: https://www.welivesecurity.com/2019/09/12/vulnerability-instagram-private-information/.

[94] Hodges, J., Jackson, C., and Barth, A. HTTP strict transport security (HSTS). *RFC* 6797 (2012). DOI: 10.17487/RFC6797.

[95] Holz, T. and Oprea, A. *IEEE S&P' 21 Program Committee Statement Regarding The "Hypocrite Commits" Paper*. 2021. URL: https://www.ieee-security.org/TC/SP2021/downloads/2021_PC_Statement.pdf (visited on 02/05/2026).

[96] HTML Living Standard. *Channel messaging*. URL: https://html.spec.whatwg.org/multipage/web-messaging.html%5C#channel-messaging (visited on 03/19/2024).

[97] HTTP Archive. *How do I use BigQuery to write custom queries over the data?* URL: https://httparchive.org/faq%5C#how-do-i-use-bigquery-to-write-custom-queries-over-the-data (visited on 12/19/2024).

[98] Huang, L., Moshchuk, A., Wang, H. J., Schecter, S., and Jackson, C. Clickjacking: attacks and defenses. In: *21st USENIX Security Symposium (USENIX Security 12)*. 2012.

[99] ICWSM. *Call for Papers 2026*. 2026. URL: https://icwsm.org/2026/submit.html%5C#paper-checklist (visited on 01/08/2026).

[100] IEEE. *IEEE Symposium on Security and Privacy 2022 – Call for Papers*. 2021. URL: https://www.ieee-security.org/TC/SP2022/cfpapers.html (visited on 01/08/2026).

[101] Inglesant, P. G. and Sasse, M. A. The True Cost of Unusable Password Policies: Password Use in the Wild. In: *ACM Conference on Human Factors in Computing Systems*. 2010.

[102] Insight EU Monitoring. *Dare More Progress: Unofficial English Translation of German Coalition Agreement Added*. 2022. URL: https://portal.ieu-monitoring.com/editorial/dare-more-progress-agreement-of-germanys-new-coalition-now-online/363687 (visited on 12/19/2024).

[103] Internet Archive. *Warcprox - WARC writing MITM HTTP/S proxy*. https://github.com/internetarchive/warcprox.

[104] Invernizzi, L., Thomas, K., Kapravelos, A., Comanescu, O., Picod, J.-M., and Bursztein, E. Cloak of Visibility: Detecting When Machines Browse a Different Web. In: *2016 IEEE Symposium on Security and Privacy (SP)*. 2016. DOI: 10.1109/SP.2016.50.

[105] Ioannidis, J. P. Why most published research findings are false. *PLoS medicine* 2, 8 (2005). DOI: 10.1371/journal.pmed.0020124.

[106] Iqbal, U., Snyder, P., Zhu, S., Livshits, B., Qian, Z., and Shafiq, Z. Adgraph: a graph-based approach to ad and tracker blocking. In: *2020 IEEE Symposium on Security and Privacy (SP)*. 2020. DOI: 10.1109/SP40000.2020.00005.

[107] Iqbal, U., Wolfe, C., Nguyen, C., Englehardt, S., and Shafiq, Z. Khaleesi: breaker of advertising and tracking request chains. In: *31st USENIX Security Symposium (USENIX Security 22)*. 2022.

[108] Jabiyev, B., Sprecher, S., Onarlioglu, K., and Kirda, E. T-Reqs: HTTP Request Smuggling with Differential Fuzzing. In: *Proceedings of the 2021 on ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 2021. DOI: 10.1145/3460120.3485384.

[109] Jenkins, R. Moral imperialism. In: *Encyclopedia of Global Justice*. Ed. by Chatterjee, D. K. Springer, 2011.

[110] Jones, B., Ensafi, R., Feamster, N., Paxson, V., and Weaver, N. Ethical Concerns for Censorship Measurement. In: *Proceedings of the 2015 ACM SIGCOMM Workshop on Ethics in Networked Systems Research*. NS Ethics '15. Association for Computing Machinery, 2015. DOI: 10.1145/2793013.2793015.

[111] Jueckstock, J. and Kapravelos, A. VisibleV8: In-browser Monitoring of JavaScript in the Wild. In: *Proceedings of the ACM Internet Measurement Conference (IMC)*. 2019. DOI: 10.1145/3355369.3355599.

[112] Jueckstock, J. et al. The blind men and the internet: multi-vantage point web measurements. *CoRR* (2019).

[113] Jueckstock, J. et al. Towards Realistic and ReproducibleWeb Crawl Measurements. In: *Proceedings of the Web Conference 2021*. 2021. DOI: 10.1145/3442381.3450050.

[114] Kant, I. *Groundwork of the Metaphysics of Morals: A German-English Edition*. Cambridge University Press, 2011.

[115] Karami, S., Ilia, P., and Polakis, J. Awakening the web's sleeper agents: misusing service workers for privacy leakage. In: *Proceedings of 2021 Network and Distributed System Security Symposium (NDSS)*. 2021. DOI: 10.14722/ndss.2021.23104.

[116] Khodayari, S. and Pellegrino, G. It's (DOM) Clobbering Time: Attack Techniques, Prevalence, and Defenses. In: *44th IEEE Symposium on Security and Privacy (SP)*. 2023. DOI: 10.1109/SP46215.2023.10179403.

[117] Kilian, M. and Schultz, U. Germany: Resistance and Reactions to Demands of Modernisation. In: *Lawyers in 21st-Century Societies*. Ed. by Abel, R. L., Hammerslev, O., Sommerlad, H., and Schultz, U. Hart, 2020.

[118] Kim, S., Kim, Y. M., Hur, J., Song, S., Lee, G., and Lee, B. FuzzOrigin: Detecting UXSS vulnerabilities in browsers through origin fuzzing. In: *31st usenix security symposium (usenix security 22)*. 2022.

[119] kingthorin and zbraiterman. *SQL Injection*. 2022. URL: https://owasp.org/www-community/attacks/SQL_Injection (visited on 02/12/2026).

[120] Kirchner, R., Möller, J., Musch, M., Klein, D., Rieck, K., and Johns, M. Dancer in the Dark: Synthesizing and Evaluating Polyglots for Blind Cross-Site Scripting. In: *33rd USENIX Security Symposium (USENIX Security 24)*. 2024.

[121] Kirchner, R., Tsoukaladelis, C., Johns, M., and Nikiforakis, N. The power to never be wrong: evasions and anachronistic attacks against web archives. In: *Proceedings of the 2025 on ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 2025. DOI: 10.1145/3719027.3765051.

[122] Kirda, E., Kruegel, C., Vigna, G., and Jovanovic, N. Noxes: a client-side solution for mitigating cross-site scripting attacks. In: *Proceedings of the 2006 ACM symposium on Applied computing*. 2006. DOI: 10.1145/1141277.1141357.

[123] Klein, D., Barber, T., Bensalim, S., Stock, B., and Johns, M. Hand Sanitizers in the Wild: A Large-scale Study of Custom JavaScript Sanitizer Functions. In: *Proceedings of the IEEE European Symposium on Security and Privacy*. 2022. DOI: 10.1109/EuroSP53844.2022.00023.

[124] Klein, M., Shankar, H., Balakireva, L., and Sompel, H. V. de. The memento tracer framework: balancing quality and scalability for web archiving. In: *International Conference on Theory and Practice of Digital Libraries*. 2019. DOI: 10.1007/978-3-030-30760-8_15.

[125] Klein, R. A. et al. Investigating variation in replicability: A "many labs" replication project. *Social Psychology* 45, 3 (2014). DOI: 10.1027/1864-9335/a000178.

[126] Klemmer, J. H. et al. How Transparent is Usable Privacy and Security Research? A Meta-Study on Current Research Transparency Practices. In: *34th USENIX Security Symposium (USENIX Security 25)*. 2025.

[127] Knittel, L., Mainka, C., Niemietz, M., NoSS, D. T., and Schwenk, J. Xsinator.com: from a formal model to the automatic evaluation of cross-site leaks in web browsers. In: *Proceedings of the 2021 on ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 2021. DOI: 10.1145/3460120.3484739.

[128] Kohno, T., Acar, Y., and Loh, W. Ethical frameworks and computer security trolley problems: foundations for conversations (2023).

[129] Komanduri, S. et al. Of Passwords and People: Measuring the Effect of Password-Composition Policies. In: *ACM Conference on Human Factors in Computing Systems (CHI)*. 2011. DOI: 10.1145/1978942.1979321.

[130] Kosek, M., Schumann, L., Marx, R., Doan, T. V., and Bajpai, V. Dns privacy with speed? evaluating dns over quic and its impact on web performance. In: *Proceedings of the 22nd ACM Internet Measurement Conference*. 2022. DOI: 10.1145/3517745.3561445.

[131] Kranch, M. and Bonneau, J. Upgrading HTTPS in Mid-Air: An Empirical Study of Strict Transport Security and Key Pinning. In: *Proceedings of 2015 Network and Distributed System Security Symposium (NDSS)*. 2015. DOI: 10.14722/ndss.2015.23162.

[132] Kremer, B., Fischer, L., and Fecher, B. Wissenschaftsbarometer 2025 - Repräsentative Bevölkerungsumfrage zu Wissenschaft und Forschung in Deutschland (2025). DOI: 10.4232/1.14680.

[133] Kumar, D. et al. Security challenges in an increasingly tangled web. In: *Proceedings of the 26th International Conference on World Wide Web*. 2017. DOI: 10.1145/3038912.3052686.

[134] Laperdrix, P., Starov, O., Chen, Q., Kapravelos, A., and Nikiforakis, N. Fingerprinting in style: detecting browser extensions via injected style sheets. In: *30th USENIX Security Symposium (USENIX Security 21)*. 2021.

[135] Le Pochat, V., Van Goethem, T., Talajizadehkhoob, S., Korczyski, M., and Joosen, W. Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation. In: *Proceedings of 2019 Network and Distributed System Security Symposium (NDSS)*. 2019. DOI: 10.14722/ndss.2019.23386.

[136] Lekies, S., Stock, B., and Johns, M. 25 million flows later: large-scale detection of DOM-based XSS. In: *Proceedings of the 2013 on ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 2013. DOI: 10.1145/2508859.2516703.

[137] Lerner, A., Kohno, T., and Roesner, F. Rewriting history: changing the archived web from the present. In: *Proceedings of the 2017 on ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 2017. DOI: 10.1145/3133956.3134042.

[138] Lerner, A., Simpson, A. K., Kohno, T., and Roesner, F. Internet jones and the raiders of the lost trackers: an archaeological study of web tracking from 1996 to 2016. In: *25st USENIX Security Symposium (USENIX Security 16)*. 2016.

[139] Lie, D. and Kirda, E. *ACM CCS 2024 Call for Papers*. URL: https://www.sigsac.org/ccs/CCS2024/call-for/call-for-papers.html (visited on 09/28/2024).

[140] Littman, M. L. Collusion rings threaten the integrity of computer science research. *Communications of the ACM* 64, 6 (2021).

[141] Luo, W., Ding, X., Wu, P., Zhang, X., Shen, Q., and Wu, Z. Scriptchecker: to tame third-party script execution with task capabilities. In: *Proceedings of 2022 Network and Distributed System Security Symposium (NDSS)*. 2022. DOI: 10.14722/ndss.2022.24382.

[142] Lupia, A., Allison, D. B., Jamieson, K. H., Heimberg, J., Skipper, M., and Wolf, S. M. Trends in US public confidence in science and opportunities for progress. *Proceedings of the National Academy of Sciences* 121, 11 (2024). DOI: 10.1073/pnas.2319488121.

[143] Macnish, K. and van der Ham, J. Ethics in cybersecurity research and practice. *Technology in Society* 63 (2020).

[144] Mao, Y. Intercoder reliability techniques: holsti method. *The sage encyclopedia of communication research methods* 2 (2017).

[145] Market Research Telecast. *Data leak at Modern Solution: House search instead of bug bounty*. 2021. URL: https://marketresearchtelecast.com/data-leak-at-modern-solution-house-search-instead-of-bug-bounty/182043/ (visited on 02/19/2026).

[146] Matte, C., Bielova, N., and Santos, C. Do Cookie Banners Respect my Choice? : Measuring Legal Compliance of Banners from IAB Europe's Transparency and Consent Framework. In: *2020 IEEE Symposium on Security and Privacy (SP)*. 2020. DOI: 10.1109/SP40000.2020.00076.

[147] Mayring, P. Qualitative content analysis. *Forum: Qualitative Sozialforschung / Forum: Qualitative Social Research* (2000).

[148] McAllister, S., Kirda, E., and Kruegel, C. Leveraging User Interactions for In-Depth Testing of Web Applications. In: *Recent Advances in Intrusion Detection: 11th International Symposium*. Springer, 2008.

[149] McDonald, N., Schoenebeck, S., and Forte, A. Reliability and inter-rater reliability in qualitative research: norms and guidelines for cscw and hci practice. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019).

[150] MDN. *Quirks Mode*. URL: https://developer.mozilla.org/en-US/docs/Web/HTML/Quirks_Mode_and_Standards_Mode (visited on 06/03/2024).

[151] Meier, R., Lenders, V., and Vanbever, L. Ditto: wan traffic obfuscation at line rate. In: *Proceedings of 2022 Network and Distributed System Security Symposium (NDSS)*. 2022. DOI: 10.14722/ndss.2021.24444.

[152] Memon, G. Leet'09: 2nd usenix workshop on large-scale exploits and emergent threats (2009).

[153] Mill, J. S. *Utilitarianism, On Liberty, Considerations on Representative Government, Remarks on Bentham's Philosophy*. Everyman, 1993.

[154] Mitmproxy Project. *mitmproxy docs*. https://docs.mitmproxy.org/.

[155] Mozilla. *gecko-dev*. URL: https://github.com/mozilla/gecko-dev (visited on 01/08/2024).

[156] Mozilla Developer Network. *Cross-Origin-Embedder-Policy*. 2023. URL: https://developer.mozilla.org/en-US/docs/Web/HTTP/Headers/Cross-Origin-Embedder-Policy (visited on 12/18/2024).

[157] Mozilla Developer Network. *Cross-Origin-Opener-Policy*. 2023. URL: https://developer.mozilla.org/en-US/docs/Web/HTTP/Headers/Cross-Origin-Opener-Policy (visited on 12/18/2024).

[158] Mozilla Developer Network. *Cross-Origin-Resource-Policy*. 2023. URL: https://developer.mozilla.org/en-US/docs/Web/HTTP/Headers/Cross-Origin-Resource-Policy (visited on 12/18/2024).

[159] Mozilla Developer Network. *Origin*. 2023. URL: https://developer.mozilla.org/en-US/docs/Web/HTTP/Headers/Origin (visited on 12/18/2024).

[160] Mozilla Developer Network. *Permissions Policy*. 2023. URL: https://developer.mozilla.org/en-US/docs/Web/HTTP/Permissions_Policy (visited on 12/18/2024).

[161] Mozilla Developer Network. *Referrer-Policy*. 2023. URL: https://developer.mozilla.org/en-US/docs/Web/HTTP/Headers/Referrer-Policy (visited on 12/18/2024).

[162] Mozilla Developer Network. *Document Object Model (DOM)*. 2026. URL: https://developer.mozilla.org/en-US/docs/Web/API/Document_Object_Model.

[163] Mozilla Developer Network. *Mixed content*. URL: https://developer.mozilla.org/en-US/docs/Web/Security/Defenses/Mixed_content (visited on 12/18/2024).

[164] Mozilla Developer Network. *Same-origin policy*. URL: https://developer.mozilla.org/en-US/docs/Web/Security/Defenses/Same-origin_policy (visited on 12/18/2024).

[165] Murphy, J., Hashim, N. H., and O'Connor, P. Take me back: validating the wayback machine. *J. Comput. Mediat. Commun.* 13, 1 (2007). DOI: 10.1111/j.1083-6101.2007.00386.x.

[166] Musch, M. and Johns, M. U Can't Debug This: Detecting JavaScript Anti-Debugging Techniques in the Wild. In: *30th USENIX Security Symposium (USENIX Security 21)*. 2021.

[167] Narayan, S. et al. Retrofitting fine grain isolation in the firefox renderer. In: *29th USENIX Security Symposium (USENIX Security 20)*. 2020.

[168] National Academies of Sciences. *Reproducibility and Replicability in Science*. National Academies Press, 2019.

[169] NeurIPS. *NeurIPS Paper Checklist Guidelines*. 2025. URL: https://neurips.cc/public/guides/PaperChecklist (visited on 01/13/2026).

[170] Nikiforakis, N. et al. You are what you include: large-scale evaluation of remote javascript inclusions. In: *Proceedings of the 2012 on ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 2012. DOI: 10.1145/2382196.2382274.

[171] *NVD - CVE-2023-4836*. URL: https://nvd.nist.gov/vuln/detail/CVE-2023-4836 (visited on 11/05/2024).

[172] Olszewski, D. et al. "Get in Researchers; We're Measuring Reproducibility": A Reproducibility Study of Machine Learning Papers in Tier 1 Security Conferences. In: *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2023. DOI: 10.1145/3576915.3623130.

[173] Osborne, C. *Urlscan.io API unwittingly leaks sensitive URLs, data*. PortSwigger. 2022. URL: https://portswigger.net/daily-swig/urlscan-io-api-unwittingly-leaks-sensitive-urls-data.

[174] OWASP. *Cross Site Scripting (XSS)*. 2023. URL: https://owasp.org/www-community/attacks/xss/ (visited on 02/12/2026).

[175] OWASP. *Path Traversal*. 2023. URL: https://owasp.org/www-community/attacks/Path%5C_Traversal (visited on 02/12/2026).

[176] OWASP. *Testing for Insecure Direct Object References*. 2023. URL: https://owasp.org/www-project-web-security-testing-guide/latest/4-Web%5C_Application%5C_Security%5C_Testing/05-Authorization%5C_Testing/04-Testing%5C_for%5C_Insecure%5C_Direct%5C_Object%5C_References (visited on 02/12/2026).

[177] OWASP. *OWASP Top 10 2025*. 2025. URL: https://owasp.org/Top10/2025/ (visited on 02/12/2026).

[178] OWASP. *Clickjacking*. 2026. URL: https://owasp.org/www-community/attacks/Clickjacking (visited on 02/12/2026).

[179] OWASP. *Cross-site leaks Cheat Sheet*. 2026. URL: https://cheatsheetseries.owasp.org/cheatsheets/XS_Leaks_Cheat_Sheet.html (visited on 02/12/2026).

[180] OWASP. *DOM Clobbering*. 2026. URL: https://domclob.xyz/domc_wiki (visited on 02/12/2026).

[181] Partridge, C. and Allman, M. Ethical considerations in network measurement papers. *Communications of the ACM* 59, 10 (2016).

[182] Pauley, E. and McDaniel, P. Understanding the Ethical Frameworks of Internet Measurement Studies. In: *Workshop on Ethics in Computer Security (EthiCS)*. 2023. DOI: 10.14722/ethics.2023.239547.

[183] Pavur, J., Strohmeier, M., Lenders, V., and Martinovic, I. Qpep: an actionable approach to secure and performant broadband from geostationary orbit. In: *Proceedings of 2021 Network and Distributed System Security Symposium (NDSS)*. 2021. DOI: 10.14722/ndss.2021.24074.

[184] Payer, M. *An overview of the Systems Circus*. 2025. URL: https://nebelwelt.net/pubstats/stat-tot.png (visited on 01/08/2026).

[185] Pellegrino, G., Johns, M., Koch, S., Backes, M., and Rossow, C. Deemon: Detecting CSRF with Dynamic Analysis and Property Graphs. In: *Proceedings of the 2017 on ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2017. DOI: 10.1145/3133956.3133959.

[186] Pletinckx, S., Borgolte, K., and Fiebig, T. Out of Sight, Out of Mind: Detecting Orphaned Web Pages at Internet-Scale. In: *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. 2021. DOI: 10.1145/3460120.3485367.

[187] Pöpper, C. and Okhravi, H. *NDSS Symposium 2025 Call for Papers*. URL: https://www.ndss-symposium.org/ndss2025/submisions/call-for-papers/ (visited on 09/28/2024).

[188] Ramulu, H. S., Schmitt, H., Rerich, B., Rachel Gonzalez Rodriguez, Kohno, Tadayoshi, and Acar, Yasemin. Ethics in Security Research: A Data-Driven Assessment of the Past, the Present, and the Possible Future. In: *Proceedings of the 2025 on ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 2025. DOI: 10.1145/3719027.3765071.

[189] Rautenstrauch, J., Mitkov, M., Helbrecht, T., Hetterich, L., and Stock, B. To auth or not to auth? a comparative analysis of the pre-and post-login security landscape. In: *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2024. DOI: 10.1109/SP54263.2024.00094.

[190] Rautenstrauch, J., Pellegrino, G., and Stock, B. The Leaky Web: Automated Discovery of Cross-Site Information Leaks in Browsers and the Web. In: *2023 IEEE Symposium on Security and Privacy (SP)*. 2023. DOI: 10.1109/SP46215.2023.10179311.

[191] Rennhard, M., Kushnir, M., Favre, O., Esposito, D., and Zahnd, V. Automating the Detection of Access Control Vulnerabilities in Web Applications. *SN Computer Science* (2022).

[192] Resnik, D. B. and Shamoo, A. E. Reproducibility and Research Integrity. *Accountability in research* 24, 2 (2017). DOI: 10.1080/08989621.2016.1257387.

[193] Roberts, R., Goldschlag, Y., Walter, R., Chung, T., Mislove, A., and Levin, D. You Are Who You Appear to Be: A Longitudinal Study of Domain Impersonation in TLS Certificates. In: *Proceedings of the 2019 on ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2019. DOI: 10.1145/3319535.3363188.

[194] Rogaway, P. The Moral Character of Cryptographic Work (2016).

[195] Rosman, T., Bosnjak, M., Silber, H., KoSSmann, J., and Heycke, T. Open science and public trust in science: Results from two studies. *Public Understanding of Science* 31, 8 (2022). DOI: 10.1177/09636625221100686.

[196] Rossi, P. H. Vignette analysis: uncovering the normative structure of complex judgments. In: *Qualitative and quantitative social research*. Ed. by Merton, R. K., Coleman, J. S., and Rossi, P. H. Free Press, 1979.

[197] Roth, S., Barron, T., Calzavara, S., Nikiforakis, N., and Stock, B. Complex Security Policy? A Longitudinal Analysis of Deployed Content Security Policies. In: *Proceedings of 2020 Network and Distributed System Security Symposium (NDSS)*. 2020. DOI: 10.14722/ndss.2020.23046.

[198] Roth, S., Calzavara, S., Wilhelm, M., Rabitti, A., and Stock, B. The Security Lottery: Measuring Client-Side Web Security Inconsistencies. In: *31st USENIX Security Symposium (USENIX Security 22)*. 2022.

[199] Roulston, K. and Choi, M. Qualitative interviews. In: *The SAGE Handbook of Qualitative Data Collection*. Ed. by Flick, U. Sage Publications, 2018.

[200] Ruth, K., Kumar, D., Wang, B., Valenta, L., and Durumeric, Z. Toppling top lists: Evaluating the accuracy of popular website lists. In: *Proceedings of the 22nd ACM Internet Measurement Conference*. 2022. DOI: 10.1145/3517745.3561444.

[201] Ruth, K. et al. A World Wide View of Browsing the World Wide Web. In: *Proceedings of the 22nd ACM Internet Measurement Conference*. 2022. DOI: 10.1145/3517745.3561418.

[202] Saltelli, A. and Funtowicz, S. What is science's crisis really about? *Futures*. Post-Normal Science in Practice 91 (2017). DOI: 10.1016/j.futures.2017.05.010.

[203] Sanchez-Rola, I., Dell'Amico, M., Balzarotti, D., Vervier, P.-A., and Bilge, L. Journey to the center of the cookie ecosystem: unraveling actors' roles and relationships. In: *2021 IEEE Symposium on Security and Privacy (SP)*. 2021. DOI: 10.1109/SP40001.2021.9796062.

[204] Scherschel, F. A. *Federal Constitutional Court rejects appeal in Modern Solution case*. 2025. URL: https://www.heise.de/en/news/Federal-Constitutional-Court-rejects-appeal-in-Modern-Solution-case-10663687.html (visited on 03/16/2026).

[205] Schloegel, M. et al. SoK: Prudent Evaluation Practices for Fuzzing. In: *2024 IEEE Symposium on Security and Privacy (SP)*. 2024. DOI: 10.1109/SP54263.2024.00137.

[206] Schreier, M. *Qualitative Content Analysis in Practice*. Sage, 2012.

[207] Schreier, M. Qualitative Content Analysis. In: *The SAGE Handbook of Qualitative Data Analysis*. Ed. by Flick, U. Sage, 2014.

[208] Selenium Project. *Selenium History*. URL: https://www.selenium.dev/history/ (visited on 06/05/2024).

[209] Senol, A., Acar, G., Humbert, M., and Borgesius, F. Z. Leaky forms: a study of email and password exfiltration before form submission. In: *31st USENIX Security Symposium (USENIX Security 22)*. 2022.

[210] Shah, H. et al. Citizen scientists investigating cookies and app gdpr compliance (2023). DOI: 10.3030/873169.

[211] Shuster, E. Fifty years later: the significance of the nuremberg code. *New England Journal of Medicine* 337, 20 (1997).

[212] Simon Pieters Glenn Adams, A. v. K. *CSS object model (css-om)*. URL: https://www.w3.org/TR/cssom-1/ (visited on 03/19/2024).

[213] Sivan-Sevilla, I., Chu, W., Liang, X., and Nissenbaum, H. *Unaccounted Privacy Violation: A Comparative Analysis of Persistent Identification of Users Across Social Contexts*. Federal Trade Commission. 2020. URL: https://www.ftc.gov/system/files/documents/public_events/1548288/privacycon-2020-ido_sivan-sevilla.pdf (visited on 02/05/2026).

[214] Small, M. W. Ethical Imperialism. In: *The SAGE Encyclopedia of Business Ethics and Society*. Ed. by Kolb, R. W. Sage, 2018. DOI: 10.4135/9781483381503.

[215] Smith, M., Snyder, P., Livshits, B., and Stefan, D. Sugarcoat: programmatically generating privacy-preserving, web-compatible resource replacements for content blocking. In: *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. 2021. DOI: 10.1145/3460120.3484578.

[216] Smith, M., Snyder, P., Haller, M., Livshits, B., Stefan, D., and Haddadi, H. Blocked or Broken? Automatically Detecting When Privacy Interventions Break Websites. *Proceedings on Privacy Enhancing Technologies* 4 (2022). DOI: 10.56553/popets-2022-0096.

[217] Snyder, P., Ansari, L., Taylor, C., and Kanich, C. Browser Feature Usage on the Modern Web. In: *Proceedings of the 2016 Internet Measurement Conference*. 2016. DOI: 10.1145/2987443.2987466.

[218] Snyder, P., Taylor, C., and Kanich, C. Most Websites Don't Need to Vibrate: A Cost-Benefit Approach to Improving Browser Security. In: *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 2017. DOI: 10.1145/3133956.3133966.

[219] So, J., Ferdman, M., and Nikiforakis, N. The More Things Change, the More They Stay the Same: Integrity of Modern JavaScript. In: *Proceedings of the ACM Web Conference 2023*. 2023. DOI: 10.1145/3543507.3583395.

[220] Solomos, K., Kristoff, J., Kanich, C., and Polakis, J. Tales of Favicons and Caches: Persistent Tracking in Modern Browsers. In: *Proceedings of 2021 Network and Distributed System Security Symposium (NDSS)*. 2021. DOI: 10.14722/ndss.2021.24202.

[221] Sompel, H. V. de, Nelson, M. L., and Sanderson, R. HTTP framework for time-based access to resource states - memento (2013). DOI: 10.17487/RFC7089.

[222] Sompel, H. V. de, Nelson, M. L., Sanderson, R., Balakireva, L., Ainsworth, S., and Shankar, H. Memento: time travel for the web. *CoRR* (2009).

[223] Soska, K. and Christin, N. Automatically detecting vulnerable websites before they turn malicious. In: *23st USENIX Security Symposium (USENIX Security 14)*. 2014.

[224] Spring, T. *Researchers: MedSec, Muddy Waters Set Bad Precedent With St. Jude Medical Short*. 2016. URL: https://threatpost.com/researchers-medsec-muddy-waters-set-bad-precedent-with-st-jude-medical-short/120266/ (visited on 01/02/2026).

[225] Stafeev, A. and Pellegrino, G. SoK: State of the Krawlers–Evaluating the Effectiveness of Crawling Algorithms for Web Security Measurements. In: *33rd USENIX Security Symposium (USENIX Security 2024)*. 2024.

[226] Staicu, C.-A. and Pradel, M. Freezing the Web: A Study of ReDoS Vulnerabilities in JavaScript-based Web Servers. In: *27st USENIX Security Symposium (USENIX Security 18)*. USENIX Association, 2018.

[227] Stamm, S., Sterne, B., and Markham, G. Reining in the web with content security policy. In: *The Web Conference*. 2010. DOI: 10.1145/1772690.1772784.

[228] StatCounter. *Browser Market Share Worldwide*. URL: https://gs.statcounter.com/browser-market-share (visited on 01/08/2024).

[229] Steffens, M., Musch, M., Johns, M., and Stock, B. Whos Hosting the Block Party? Studying Third-Party Blockage of CSP and SRI. In: *Proceedings 2021 Network and Distributed System Security Symposium (NDSS)*. 2021. DOI: 10.14722/ndss.2021.24028.

[230] Steffens, M., Musch, M., Johns, M., and Stock, B. *Open-sourced version of SMURF*. https://smurf-ndss.github.io/.

[231] Steffens, M., Rossow, C., Johns, M., and Stock, B. Don't Trust The Locals: Investigating the Prevalence of Persistent Client-Side Cross-Site Scripting in the Wild. In: *Proceedings of 2019 Network and Distributed System Security Symposium (NDSS)*. 2019. DOI: 10.14722/ndss.2019.23009.

[232] Stock, B., Johns, M., Steffens, M., and Backes, M. How the Web Tangled Itself: Uncovering the History of Client-Side Web (In) Security. In: *26th USENIX Security Symposium (USENIX Security 17)*. 2017.

[233] Stock, B., Lekies, S., Mueller, T., Spiegel, P., and Johns, M. Precise client-side protection against DOM-based cross-site scripting. In: *23rd USENIX Security Symposium (USENIX Security 14)*. USENIX Association, 2014.

[234] Stock, B., Pellegrino, G., Rossow, C., Johns, M., and Backes, M. Hey, You Have a Problem: On the Feasibility of Large-Scale Web Vulnerability Notification. In: *25st USENIX Security Symposium (USENIX Security 16)*. USENIX Association, 2016.

[235] Stock, B., Pfistner, S., Kaiser, B., Lekies, S., and Johns, M. From facepalm to brain bender: exploring client-side cross-site scripting. In: *Proceedings of the 2015 on ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 2015. DOI: 10.1145/2810103.2813625.

[236] Strauss, A. L. and Corbin, J. M. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. 3rd ed. Sage, 2008.

[237] Szurdi, J., Kocso, B., Cseh, G., Spring, J., Felegyhazi, M., and Kanich, C. The long Taile of typosquatting domain names. In: *23rd USENIX Security Symposium (USENIX Security 14)*. 2014.

[238] Tan, M., Wan, J., Zhou, Z., and Li, Z. Invisible probe: timing attacks with pcie congestion side-channel. In: *2021 IEEE Symposium on Security and Privacy (SP)*. 2021. DOI: 10.1109/SP40001.2021.00059.

[239] TanviHacks. *Mixed Content Blocking Enabled in Firefox 23!* URL: https://blog.mozilla.org/tanvi/2013/04/10/mixed-content-blocking-enabled-in-firefox-23/ (visited on 02/07/2026).

[240] The Chromium Project. *Chromium*. https://github.com/chromium/chromium.

[241] The Economist. Problems with scientific research: How Science Goes Wrong. *The Economist* (2013).

[242] The Tor Project. *Research Safety Board*. 2022. URL: https://research.torproject.org/safetyboard/ (visited on 12/22/2022).

[243] Thomas, D. R., Pastrana, S., Hutchings, A., Clayton, R., and Beresford, A. R. Ethical issues in research using datasets of illicit origin. In: *Proceedings of the 2017 Internet Measurement Conference*. 2017. DOI: 10.1145/3131365.3131389.

[244] Todd W Rice. The Historical, Ethical, and Legal Background of Human-Subjects Research. *Respiratory Care* (2008). DOI: 10.4187/respcare.08531325.

[245] United States National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research*. Vol. 1. Department of Health, Education, and Welfare, 1978.

[246] USENIX. *Guidance for Revising Ethical Considerations for USENIX Security 2026*. 2026. URL: https://docs.google.com/document/d/e/2PACX-1vTBAZ4XrPejBY6EhPYhkXAS9l0pd9Qsl2HPbW26LJJOGp7gSGeWnd1mdoTC8kZuL95PvajiU4oBS4Ew (visited on 02/03/2026).

[247] Utz, C., Amft, S., Degeling, M., Holz, T., Fahl, S., and Schaub, F. Privacy Rarely Considered: Exploring Considerations in the Adoption of Third-Party Services by Websites. *Proceedings on Privacy Enhancing Technologies* 2023, 1 (2023). DOI: 10.56553/popets-2023-0002.

[248] WebKit. *A fast, open source web browser engine*. https://webkit.org/.

[249] Webrecorder. *pywb*. https://github.com/webrecorder/pywb.

[250] Weichselbaum, L., Spagnuolo, M., Lekies, S., and Janc, A. CSP is dead, long live csp! on the insecurity of whitelists and the future of content security policy. In: *Proceedings of the 2016 on ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 2016. DOI: 10.1145/2976749.2978363.

[251] Weissbacher, M., Lauinger, T., and Robertson, W. Why is csp failing? trends and challenges in csp adoption. In: *International Workshop on Recent Advances in Intrusion Detection*. 2014. DOI: 10.1007/978-3-319-11379-1_11.

[252] Wermke, D., Wöhler, N., Klemmer, J. H., Fourné, M., Acar, Y., and Fahl, S. Committed to trust: A qualitative study on security & trust in open source software projects. In: *2022 IEEE symposium on Security and Privacy (SP)*. IEEE. 2022. DOI: 10.1109/SP46214.2022.9833686.

[253] West, M. *Block requests for suspected dangling markup*. 2017. URL: https://github.com/whatwg/fetch/pull/519 (visited on 02/17/2026).

[254] WHATWG. *HTML Living Standard*. URL: https://html.spec.whatwg.org (visited on 02/06/2026).

[255] WHATWG. *HTML Standard: Link type "preload"*. URL: https://html.spec.whatwg.org/multipage/links.html%5C#link-type-preload (visited on 06/05/2024).

[256] Witzel, A. and Reiter, H. *The Problem-centered Interview: Principles and Practice*. Sage, 2012.

[257] World Wide Web Consortium (W3C). *Upgrade Insecure Requests*. 2015. URL: https://www.w3.org/TR/upgrade-insecure-requests/ (visited on 03/07/2026).

[258] World Wide Web Consortium (W3C). *WebDriver*. 2018. URL: https://www.w3.org/TR/webdriver1 (visited on 06/05/2024).

[259] World Wide Web Consortium (W3C). *Mixed Content*. 2023. URL: https://w3c.github.io/webappsec-mixed-content (visited on 06/05/2024).

[260] World Wide Web Consortium (W3C). *Content Security Policy (CSP)*. 2024. URL: https://w3c.github.io/webappsec-csp/ (visited on 06/05/2024).

[261] WPT Contributors. *web-platform-tests documentation*. URL: https://web-platform-tests.org (visited on 06/05/2024).

[262] Wu, Q. and Lu, K. *On the Feasibility of Stealthily Introducing Vulnerabilities in Open-Source Software via Hypocrite Commits*. 2021. URL: https://github.com/QiushiWu/QiushiWu.github.io/blob/main/papers/OpenSourceInsecurity.pdf (visited on 06/05/2025).

[263] Yong, E. How Reliable Are Psychology Studies? *The Atlantic* (2015).

[264] Zagi, L., Moti, Z., and Acar, G. Referrer Policy: Implementation and Circumvention. *Proceedings on Privacy Enhancing Technologies (PETS)* (2025). DOI: 10.56553/popets-2025-0092.

[265] Zapp, M. The legitimacy of science and the populist backlash: Cross-national and longitudinal trends and determinants of attitudes toward science. *Public Understanding of Science* 31, 7 (2022). DOI: 10.1177/09636625221093897.

[266] Zhan, Z., Zhang, Z., Liang, S., Yao, F., and Koutsoukos, X. Graphics peeping unit: exploiting em side-channel information of gpus to eavesdrop on your neighbors. In: *2022 IEEE Symposium on Security and Privacy (SP)*. 2022. DOI: 10.1109/SP46214.2022.9833773.

[267] Zhang, J., Psounis, K., Haroon, M., and Shafiq, Z. Harpo: learning to subvert online behavioral advertising. In: *Proceedings of 2021 Network and Distributed System Security Symposium (NDSS)*. 2021. DOI: 10.14722/ndss.2022.23062.

[268] Zhang, Y., Liu, M., Zhang, M., Lu, C., and Duan, H. Ethics in Security Research: Visions, Reality, and Paths Forward. In: *EuroS&PW '22*. 2022. DOI: 10.1109/EuroSPW55150.2022.00064.