

Researchers' Experiences in Analyzing Privacy Policies: Challenges and Opportunities

Abraham Mhaidli
University of Michigan
mhaidli@umich.edu

Gina Herakovic
University of Michigan
gherakov@umich.edu

Shomir Wilson
Penn State University
shomir@psu.edu

Selin Fidan
University of Michigan
selfidan@umich.edu

Mukund Srinath
Penn State University
mus824@psu.edu

Florian Schaub
University of Michigan
fschaub@umich.edu

An Doan
University of Michigan
anvidoan@umich.edu

Lee Matheson
Future of Privacy Forum
lmatheson@fpf.org

ABSTRACT

Companies' privacy policies and their contents are being analyzed for many reasons, including to assess the readability, usability, and utility of privacy policies; to extract and analyze data practices of apps and websites; to assess compliance of companies with relevant laws and their own privacy policies, and to develop tools and machine learning models to summarize and read policies. Despite the importance and interest in studying privacy policies from researchers, regulators, and privacy activists, few best practices or approaches have emerged and infrastructure and tool support is scarce or scattered. In order to provide insight into how researchers study privacy policies and the challenges they face when doing so, we conducted 26 interviews with researchers from various disciplines who have conducted research on privacy policies. We provide insights on a range of challenges around policy selection, policy retrieval, and policy content analysis, as well as multiple overarching challenges researchers experienced across the research process. Based on our findings, we discuss opportunities to better facilitate privacy policy research, including research directions for methodologically advancing privacy policy analysis, potential structural changes around privacy policies, and avenues for fostering an interdisciplinary research community and maturing the field.

KEYWORDS

Privacy, privacy policies, research practice, public policy, natural language processing, machine learning, research infrastructure.

1 INTRODUCTION

Privacy policies are documents in which companies, organizations, or services describe their data practices, including how they collect, use, store, and share consumer data and personally identifiable information. They can be ways for consumers to learn how companies collect their data as well as what privacy rights they have and

how to exercise them. Their importance is such that, in some legal jurisdictions, companies that process consumer data are required to provide privacy policies.

For these reasons, privacy policies are also of interest to researchers. Some researchers have studied privacy policies to understand the data practices of certain companies and devices [38]. Other work has examined the usability of privacy policies, such as the readability or complexity of privacy documents [54]. Some scholars have further created tools that help summarize and analyze privacy policies, such as tools that can summarize privacy policy information for consumers [2, 8] or regulators [56].

Despite increased work in this area, privacy policies remain difficult for researchers to analyze and no clear best practices or methodologies for privacy policy analysis have emerged so far. Furthermore, tool and infrastructure support are relatively scarce. Our work provides insights into different ways in which researchers approach, study, and analyze privacy policies, what core challenges researchers face when studying privacy policies, and what opportunities exist for better supporting this line of research.

To that end, we conducted semi-structured interviews with 26 researchers from various disciplines who have worked with privacy policies to understand how they study privacy policies and what challenges they have faced in doing so. We found that in the absence of established best practices most participants developed their own approaches for studying privacy policies. Furthermore, we found that researchers faced numerous challenges when studying privacy policies, including: challenges related to *policy selection*, i.e., what policies to study; *policy retrieval*, i.e., finding and collecting privacy policies, such as difficulties automating the collection of policies and issues with inconsistent file formats and privacy policy locations on websites; challenges related to *policy analysis*, including privacy policies' excessive use of jargon and vagueness, both of which complicate scaling up analysis; as well as *overarching challenges*, including lack of availability and maintenance of research tools, challenges with non-English policies, a lack of privacy policy-specific venues for the presentation and publication of relevant work, and difficulties in forming interdisciplinary collaborations.

Our findings provide insights on opportunities for better supporting privacy policy research in four key areas: opportunities for research that will help mature the field; opportunities for research

This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license visit <https://creativecommons.org/licenses/by/4.0/> or send a letter to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.



Proceedings on Privacy Enhancing Technologies 2023(4), 287–305
© 2023 Copyright held by the owner/author(s).
<https://doi.org/10.56553/popets-2023-0111>

tool development; opportunities for companies and policymakers to better facilitate access to privacy policies; and opportunities for building a more robust privacy policy research community that facilitates cross-disciplinary collaboration and recognizes the technical challenges involved in privacy policy research.

Addressing the identified challenges in privacy policy research will not only improve workflows for researchers but allow for the development of tools and infrastructures that better help consumers and data subjects navigate complicated privacy policies, as well as help increase transparency and accountability of corporate privacy promises and companies' data practices, which can inform, motivate and catalyze policy-making to better safeguard consumer privacy.

2 BACKGROUND

Privacy policy analysis has received increasing interest from first readability studies in the 1990s to substantial efforts to manually and then automatically analyze policies. The analysis of privacy policies has received attention from a wide variety of disciplines, including healthcare [38], law [36], computer science [8], privacy and security [45], machine learning [35], digital entertainment [39], media and communication [17, 32], and more. Some scholars study privacy policies to determine the data practices of certain entities; for example, studying the data practices of health apps [38]. Sometimes this includes determining whether a privacy policy accurately reflects the data practices of the company by comparing a firm's actual behavior to the policy's text [10, 42, 55, 56]. Other scholars study privacy policies to examine their usability and readability. Privacy policies are notoriously lengthy and difficult to read [16, 27], and some researchers have sought to quantify policies' vagueness [23, 25] and legibility, such as by studying policy length [3], estimating required reading comprehension [54], and measuring whether policy users can actually demonstrate comprehension [36]. Further work has examined the interplay between privacy policies and legislation, including whether selected services comply with relevant legislation such as the European Union's General Data Protection Regulation (GDPR) [29, 35, 48], and what impact particular legislation has had on privacy policies [15, 52]. Researchers have also created corpora of privacy policies, such as the annotated OPP-115 Corpus [49], the Princeton-Leuven Longitudinal Corpus [3], and the PrivaSeer corpus [44].

Natural Language Processing (NLP) researchers have contributed significantly to the study of privacy policies. A number of NLP techniques have been applied to extract key information from or to summarize privacy policies [22, 53]. Question answering has been a popular approach to simplifying privacy policies. PriBot [18] enables free-form queries on privacy policies. Ravichander et al. [34] created the PrivacyQA corpus by asking crowd workers to ask privacy questions about mobile apps to help improve question-answering approaches. Similarly Ahmad et al. [1] created the PolicyQA dataset to help retrieve relevant text spans from privacy policies. Wilson et al. [49] used expert annotations from legal scholars to categorize privacy policies into their constituent data practices. Opt-Out Easy automatically extracts opt-out links from privacy policies [7]. PrivacyCheckV2 [31] and PrivacyCheckV3 [30] not only summarize privacy policies, but also run competitor analyses that highlight the top privacy-preserving alternatives to a company in a particular

sector. PrivaSeer is a privacy policy search engine where policies can be filtered based on a number of facets [43]. Other projects have compared privacy policies across multiple languages [6]. The Usable Privacy Policy Project (UPPP)¹ has played an important role in catalyzing research in this area by creating policy corpora, analysis techniques, and tools that allow users to better understand and control their privacy, for example, by answering questions about privacy policies that users care about [40, 50].

However, while many have studied privacy policies, *how* researchers conduct this research and what challenges they experience in the process is less understood. It appears that there are few consistent processes or best practices for the analysis of privacy policies. We can glean a rough understanding of privacy policy researchers' methodologies by reviewing existing work. Scholars use a wide array of methods to analyze privacy policies, including manually annotating and coding policies [41], using crowd workers to read and evaluate policies [50], and deploying online readability calculators or NLP and ML techniques to examine privacy policies and extract useful information from them [7, 14, 24]. We can similarly infer some of the challenges researchers likely face when studying privacy policies. For example, privacy policies are notoriously vague and difficult to interpret by users [16, 27]; there has been work showing that even experts might disagree on the interpretation of privacy policies [36]. The abundance of papers seeking to automate or expedite the analysis of privacy policies (e.g., see [4, 18, 51]) implies that privacy policy analysis is tedious and time-consuming. Other papers explicitly identify particular problems in the privacy space. For example, Harkous et al. [18] highlight a "significant scalability hurdle" in the study of privacy policies. These papers imply that the length, language, and complexity of privacy policies is a major challenge to effective analysis; and as a consequence, studying them is tedious, time-consuming, and difficult to automate and scale. However, while we can make some assumptions about how researchers study privacy policies, and what challenges they face in doing so, a direct analysis of how researchers study privacy policies is overdue. Firsthand accounts from researchers can illuminate challenges in the research process as well as opportunities for overcoming these challenges and improving the field [26].

3 METHOD

In our study, we explore how researchers study privacy policies to gain insights into what challenges they face when doing so. To accomplish this, we conducted semi-structured interviews with 26 researchers from various disciplines who have studied privacy policies at some point in their research. Interviews with researchers can be effective tools for unearthing challenges and methodologies [26], since they allow in-depth probing into participants' approaches as well as the rationale for those approaches. Our study was reviewed and deemed exempt from oversight by the University of Michigan's Institutional Review Board (IRB).

3.1 Recruitment and Outreach

To find researchers with experience studying privacy policies, we performed searches on popular scholarly search engines (Google

¹<https://usableprivacy.org/>

Scholar,² Semantic Scholar,³ and WorldWideScience⁴ for research articles and papers published in the last five years (2017–2022) that studied privacy policies. We used the search term “*privacy policy*” OR “*privacy policies*” and selected articles that mention these terms in their abstract. We excluded papers that were (co-)authored by members of our research team. For each search engine, we selected the first 100 articles that matched these criteria. The research team manually validated that these papers indeed involved the analysis of privacy policies as part of their methodology. After this step (and after removing duplicates), we had a list of 125 papers that studied privacy policies.

We reached out to the authors of these papers in batches of 20. First, we randomly selected 20 papers from our sample. For each paper, we selected the paper’s designated contact author, or if none were given, the first author of that paper. We sent a recruitment email inviting them to participate in our interview study, with a scheduled follow-up message one week later. If the authors declined the interview request or did not respond after the second email, we removed that paper from our sample. After all papers in a batch had either been contacted or removed, we moved to the next batch until we reached saturation in the interview data.

After the first 15 interviews were completed, we observed that our sample of responding participants skewed towards US and computer science participants. To stratify our sample for geographical and disciplinary diversity, in subsequent rounds we selected papers that were written by authors in departments outside of the US and/or in non-CS departments. We kept interviewing participants until data saturation was reached, that is, we repeatedly observed the same themes in the interviews and no new themes emerged [28].

3.2 Interview Protocol

We conducted semi-structured interviews with our participants to learn about their experiences, practices, and challenges in working with privacy policies. In the first part of the interview, we asked general questions about the participant including their research background, their experience with studying privacy policies, and how the analysis of privacy policies fit into their work.

We then inquired about a specific paper or study they had conducted to understand the process by which the participant had studied privacy policies—this helped us anchor the interview conversation on a specific project, for which we could probe deeply about the rationale for their topical and methodological decisions, while also encouraging participants to discuss how this particular project was exemplary or different from other projects and their general research approach regarding privacy policies. For example, if they had studied a privacy policy using a certain method, why did they pick that method? If they studied a certain number of policies, how did they arrive at that particular number?

We then asked specifically about what challenges, if any, our participants had faced when studying privacy policies. Lastly, we asked what tools or infrastructures participants would find helpful

in studying privacy policies, as well as asked for participants’ reactions and thoughts on a set of proposed tools and directions. The full interview protocol can be found in Appendix A.

Interviews were conducted remotely via Zoom videoconferencing. The interviews lasted around 1 hour, with a median time of 55:22, a mean of 54:39, a min of 35:30, and a max of 1:07:42. Each interview was recorded using Zoom’s recording feature, to help in subsequent data analysis. Interview participants received \$25 USD compensation in the form of either a mailed check or an Amazon gift card. One participant was interviewed via email following their request.

3.3 Participant Information

We interviewed a total of 26 researchers. To protect participants’ anonymity, some of whom are prominent researchers in the field, we report participant demographics in the aggregate, rather than tying each participant ID to their demographic information.

We asked participants for their preferred gender pronouns. 12 participants used he/him pronouns; nine used she/her pronouns; one used they/them pronouns; and four participants chose not to disclose their preferred pronouns. Most of our participants (21) worked in academia, and the rest in industry. Our participants came from various disciplines as reflected by their departments, including computer science and computing (15), health (three), communication (two), policy (two), education (one), business (one), sociology (one) and political science (one). Some participants had appointments in multiple departments or worked in departments that covered two or more of the above areas. Out of the fourteen participants working in computing-related fields, ten worked in machine learning and/or natural language processing. Our participants were diverse in terms of roles and level of seniority. Eight participants were assistant professors, two were associate professors, one was a lecturer, two were research fellows, and five were Ph.D. students. Other job titles included research scientist (two) and research director (one). A further five participants worked in an industry setting for software companies. This included two privacy engineers, one software engineer, one software developer, and one applied scientist. Sixteen of our participants worked in the US, three in Germany, two in the UK, two in China, one in Australia, one in Canada, and one in Morocco.

Our participants were also diverse in terms of the research they conducted with privacy policies. Participants studied privacy policies to understand the data practices of companies (17), including comparing what a company’s policy stated versus its actual data practices (five); analyzing compliance with legislation (ten); studying the usability of privacy policies (six); developing new tools or frameworks to analyze privacy policies, including training machine learning models (seven); as well as discourse analysis of privacy policies (one). Most participants had studied privacy policies for several years, but five participants had only recently entered the field (three years or fewer). Similarly, while most participants worked on numerous projects involving privacy policies, nine participants had worked on relatively few projects, with six participants only working on one project involving privacy policy analysis. 18 participants were still studying and researching privacy policies; eight

²<https://scholar.google.com/>

³<https://www.semanticscholar.org/>

⁴<https://worldwidescience.org/>

participants no longer worked with privacy policies. Lastly, participants differed in terms of how prominent a part privacy policy analysis was of their research agenda. Out of those who still worked on privacy policies, for three participants privacy policy analysis formed a large part of their work; for eight it formed some part of their work; and for seven it formed only a small part of their research agenda.

3.4 Data Analysis

To analyze our data, we first obtained interview transcripts using Zoom’s built-in transcription service. We reviewed the transcripts and corrected them for accuracy. Two of the authors subsequently coded the interview transcripts using thematic coding [28] as follows. The coders first jointly created an initial version of the codebook to capture key themes from the interviews. They then iterated on the codebook by analyzing interviews and modifying the codebook until the codebook captured all major themes. The authors then independently coded interviews and compared inter-rater reliability (IRR) until a high IRR was reached (Cohen’s Kappa > 0.75). Once this occurred, the authors each independently (re-)coded 13 of the interviews. The final version of the codebook contains 52 codes and can be found in Appendix B.

3.5 Limitations

Our sample size (26) is typical for qualitative interview studies. Yet, the qualitative nature of the study means our findings may not be representative of all researchers’ experiences with privacy policy research. The goal of our study was not generalizability: rather, it was to identify challenges within privacy policy research. We acknowledge that the findings of this study cannot speak to how all privacy policy research works. We are not saying that every researcher experiences all challenges discovered here, and we cannot say how prevalent each of the challenges is—but our findings shed insights into the challenges that researchers face. There might also be other challenges that exist and are not captured here, but we were diligent in stratifying our sample both geographically and in terms of experience so we are confident that our findings are reasonably representative of challenges that occur when analyzing privacy policies.

Furthermore, we only capture the opinions of those who have successfully published a paper that has studied privacy policies. This means that our study does not reflect the voices of those who have not (yet) published research with privacy policies and those who found studying privacy policies so challenging that they did not work with them. This was a necessary choice to more efficiently recruit participants we could guarantee had experience studying privacy policies. Future work can and should cover these missing perspectives.

4 FINDINGS

Our findings surface numerous challenges that researchers face when studying privacy policies and the ways that researchers work to overcome these challenges. We find that there is a lack of best practices researchers can follow; instead, our participants took it upon themselves to discover workarounds to their problems. This

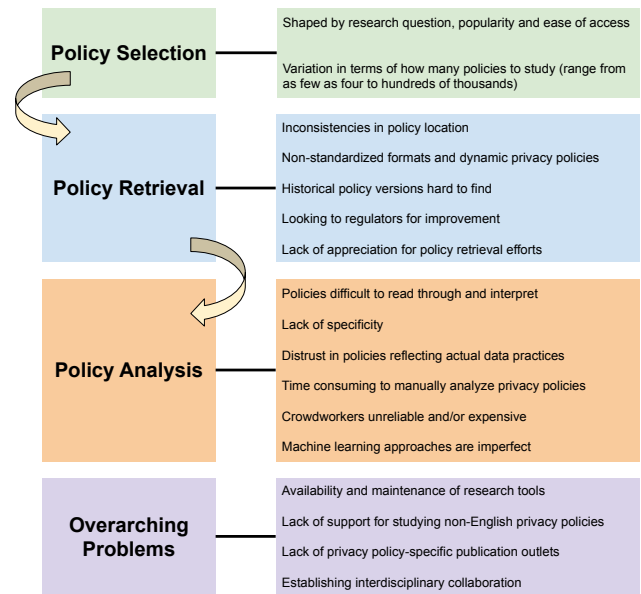


Figure 1: An overview of the steps involved in privacy policy analysis and challenges found in each step.

findings section thus reports on key challenges researchers faced and factors they took into account when analyzing privacy policies.

We organize our findings around three key steps of studying privacy policies (see Figure 1): (1) policy selection, i.e., deciding what privacy policies to study; (2) policy retrieval, i.e., searching, finding, and retrieving relevant privacy policy documents; (3) policy analysis, i.e., analyzing the retrieved privacy policies, which ranges from manually annotating policies to applying machine learning techniques to extract information from the policy. We highlight what each step consists of, how our participants approach these steps, and crucially, what key challenges participants reported facing in each step. We conclude the findings section by discussing overarching challenges for privacy policy analysis that emerged from our interviews, including challenges with sharing and maintaining research tools, a lack of privacy policy-specific publication venues, difficulties establishing interdisciplinary collaboration, and a lack of support for studying non-English privacy policies.

4.1 Policy Selection

The first step in studying privacy policies is deciding what policies to analyze. Participants took into account factors such as where or how a privacy policy might be available (e.g., websites, mobile apps, IoT devices, or some combination thereof) and the business sector the policies pertained to (e.g., health apps, banking sector, social media sites) to determine selection criteria. Not surprisingly, what privacy policies researchers chose to analyze was highly dependent on their research questions. However, the popularity of services (e.g., website ranking) and ease of access heavily shaped participants’ considerations.

4.1.1 Popularity. Many participants based their decision of which privacy policies to study on websites’ popularity, e.g., relying on the

Alexa ranking⁵ or the Tranco list.⁶ Many participants would pick the top-ranked websites in terms of traffic and study their privacy policies. Other participants had minimum popularity thresholds as part of their selection criteria; in the case of mobile apps, for example, several participants mentioned only selecting apps that had a minimum number of downloads (e.g., at least 50 downloads).

The choice of popular apps was based on the potential impact the findings could have. P25 put it as follows: *"So the most popular apps are the ones which are used the most by the consumers, the users, end users, and so if we study about the popular apps, so this subset of apps are in a way reaching out to most of the audiences. And so, whatever analysis and research we do on this subset of apps, the findings kind of impact a lot of users, and so [...] that gives a strong intuition to pick an app that's used by a lot of people."* P23 echoed a similar sentiment: *"if you're going to assess privacy policies for an app that it's not really used, [...] you can't anticipate your results having some sort of impact because it's not affecting anyone."*

4.1.2 Ease of access. Another selection criterion for researchers was ease of access. Researchers made decisions about what policies to study based on factors such as how easy the policies were to access or familiarity with the materials and technologies required. For example, P17 explained their decision to study the privacy policies of iOS apps as *"mostly just a function of I have an Apple product."* Other participants chose to study Android privacy policies either because the Android store was easier to crawl through, or, in studies that involved comparing a privacy policy's described data practices against app behavior, because there were more tools for analyzing the code of Android apps: *"most papers will only focus on the Android apps since there are many tools related to analyzing privacy of Android apps. We also want to analyze the privacy-related behaviors of the iOS apps, but with iPhone, it is very challenging to perform static analysis on iOS apps"* (P11).

In a similar vein, many participants made use of existing datasets of privacy policies, most notably the OPP-115 corpus [49]. Using such datasets saved participants a significant amount of time and labor. Per P25: *"it would have taken considerable amount of time just to annotate privacy policies yourselves, and then annotate a significant number of the privacy policies, so yeah, [using OPP-115] really helped with, you know, saving over time."* Other factors that fell under the theme of ease of access included whether a policy had been cited or used in prior work, whether an approach or framework had been used by other members of the research team or research group, and in the case of PhD students, their advisor's preferred approach or line of research. In explaining why they analyzed the websites they did, P16 said *"if I'm being honest [...] my advisor was doing work in that, and when I came in, it was more like 'hey we want to work on these,' and then I just like continued the work."*

4.1.3 Number of Policies Driven by Methodology. Deciding what policies to study also means deciding how many policies to study. Here we saw a wide range across participants. Some participants worked on studies that analyzed as few as four privacy policies, and some as many as tens and even hundreds of thousands. One

reason for the variation stemmed from the research question potentially limiting how many privacy policies one could find: if a participant was studying the privacy policies of a highly specific type of service (e.g., mobile apps for fertility/period tracking), the number of services—and hence policies to analyze—would be limited. Conversely, participants studying a broader range of privacy policies had more privacy policies from which to choose and often needed to decide on sampling criteria.

The number of policies studied was also influenced by the researcher's methodological approach. Authors who used manual analysis and qualitative methods generally studied fewer privacy policies (usually in the tens of privacy policies) than those who were trying to build machine learning classifiers or carry out large-scale statistical analysis (in the hundreds or possibly thousands of policies). Methodological choice meant that there were practical limitations as to how many policies one could study (it is much more work to manually read through and annotate 100 policies than it is to calculate readability scores for 100 privacy policies) as well as conventions of each method determining how many policies are 'enough.' P1 annotated a few hundred privacy policies to train a machine learning classifier; in explaining why they studied this number of privacy policies, P1 said: *"I had the sense that [for] the techniques that I used, it was kind of sufficient to get reasonably good results. Adding another hundred policies, for example, would maybe slightly improve but not really much. I mean this is more an anecdotal finding, it's not like I did really hard analysis on that, but I still think it's true. If you get to 100 maybe 200 [policies], then that's kind of sufficient for the things you want to do with machine learning analysis."*

In terms of how participants decided how many privacy policies to study, we saw two main approaches. Some participants took an ad-hoc approach, starting with knowing *what* privacy policies to analyze but without a clear idea of *how many*. The scale was decided as the project progressed, influenced by factors such as how many privacy policies the authors could find, practical limitations over how many could be gathered, and particularly for qualitative studies, whether data saturation was reached (i.e., no new themes emerged from analyzing additional privacy policies) and if it made sense to look for and analyze additional policies. Other participants were more systematic, starting with a fixed number of privacy policies in mind—this was often the case for participants developing machine learning models, who had an idea of how many policies would be needed to create an adequate training dataset.

4.2 Policy Retrieval

Once selection criteria had been established, the next step for researchers was finding and retrieving respective policies. Methods for finding privacy policies included going directly to companies' websites, using existing datasets (e.g., the OPP-115 corpus) and ranking lists (e.g., Alexa's top websites), searching with popular search engines, and in the case of mobile apps retrieving privacy policy URLs from app listings in Apple and Google's app stores. Depending on the scale of the analysis, some researchers located policies manually while others automated the process, e.g., by building crawlers that look through websites or app stores and collect privacy policies.

⁵The Alexa ranking was an Amazon service that ranked websites based on their web traffic; it was discontinued in 2022.

⁶<https://tranco-list.eu/>

4.2.1 Inconsistencies in Policy Location. While some participants mentioned that finding privacy policies was a fairly easy process, others complained that finding privacy policies was time-consuming and difficult. Key difficulties included broken privacy policy links or pages, or pages that linked to a general website instead of the specific privacy policy. This is further complicated by the lack of a standardized location for privacy policies across websites and apps. For any given service, the privacy policy could be in multiple documents or locations, or not exist at all. This makes manually searching for documents difficult and using an automated crawler to find privacy policies particularly challenging. One participant addressed this challenge by building their own machine learning classifier to crawl webpages and automatically identify whether the page being crawled was a privacy policy or not.

4.2.2 Non-standardized Formats and Dynamic Privacy Policies. Continuing along this theme of inconsistencies, another challenge participants faced was inconsistent privacy policy file formats, which made automating the collection of policies with web scrapers difficult. Participants mentioned that most websites posted privacy policies as HTML documents, but other websites posted policies as PDF documents or even JPEG files. Furthermore, participants mentioned that websites' privacy policy pages sometimes dynamically loaded content via JavaScript. These inconsistencies made collecting and parsing privacy policies difficult and caused significant technical difficulties and required extra tooling, such as having to use headless browsers to retrieve policy documents in their entirety, especially when trying to collect an extensive number of policies: *"Most [privacy policies] are JavaScript based, so you can't just make a HTTP GET request and get those policies because they're not pure text. So because they are JavaScript based, you will need to run Selenium browsers that will open up, and then you'll have to download these policies, now because of that, and you know how you might know that Selenium browsers are RAM hungry, so what happens is the Docker containers continue to get bogged down, and then the container just dies"* (P20).

4.2.3 Historical Policy Versions Hard to Find. Participants also mentioned that historical versions of existing policies were difficult to find. Finding historical versions of any document can be challenging, but this problem can be particularly salient for privacy policies, given that companies may frequently update their privacy policies and replace prior versions, there can be regulatory changes (which in turn prompt companies to change their privacy policies to adhere to these changes), and privacy policies may be lost in cases of bankruptcy or company acquisitions. Participants sometimes used websites like the Internet Archive's Wayback Machine⁷ to find historical versions of privacy policies, but oftentimes these were tedious to use: *"It's really challenging to find historical policies, so if you are interested in looking at kind of the evolution of a company's policy over time, often those previous versions are no longer posted on their site, you have to go to something to like the Internet archive or see if anyone else has downloaded and saved them as a database somewhere, but that gets a little bit tedious, especially if you want to do it for more than just a few companies, then you're sort of looking*

at a task in front of you, and it's not even possible to find them" (P12). Additionally, some participants mentioned that tools like the Wayback Machine were not reliable (*"sometimes [they] don't work,"* P21) and did not necessarily include a complete history of all versions of a website's privacy policy. Thus, participants mentioned that certain types of analysis were difficult to conduct, such as historical analysis or studying the effects of new legislation by comparing policies before and after the legislation was enacted.

4.2.4 Looking to Regulators for Improvement. Some participants said that regulatory changes are making it easier to consistently locate policies. For instance, P17 mentioned wanting to re-do their study given *"a lot of things have changed since I did [my study], one thing that has changed is the California privacy law has changed, that exists now, it didn't exist, then. Apple has added their privacy labels, which didn't exist, then, but do exist now... also the FTC is making some serious moves on privacy as well, and so there's a chance that finding privacy policies is easier now than it was when we did our data collection in 2018."*

Participants further noted that regulators should do more to make privacy policies more accessible. When asked about what resources or infrastructure would help facilitate privacy policy analysis, many participants expressed a desire for regulators to specify the location where privacy policies have to be posted, similar to how mobile app stores now require app listings to include a link to the app's privacy policy. P17 said: *"If there was a law that required apps to post their privacy policy somewhere that I could get it and it was required to be up to date, that would be fantastic, and I know that seems so obvious, right, like you know the App Store can require that you post privacy policy, it may or may not be there, but I would love it if something with some teeth, just put these documents in one place, and I could go there and feel reasonably confident that the thing I got was the thing that applied."* P17 continued by suggesting that organizations' updates to privacy policies should also be better regulated, including requirements for documentation of when the policy changed and how: *"I would love it if there were more restrictions on how and when these terms can change, and of course I understand it's technology it's a rapidly evolving tool, you need to be able to adapt documents to that and yada yada yada, but I would love it if there was [...] a certain amount of reliability and predictability with not only where I found these documents, but when and how they changed."*

4.2.5 Lack of Appreciation for Policy Retrieval Efforts. Participants mentioned that there was a lack of appreciation for the 'behind-the-scenes' work spent on policy retrieval, such as searching for, filtering, and formatting privacy policies for analysis. Depending on the specific study, participants spent a substantial amount of time on these tasks, and this work was often difficult, tedious, and/or time-consuming, both when done manually as well as when researchers invested engineering effort to build out a more automated policy retrieval infrastructure. Yet, our participants felt that these efforts in policy retrieval and data cleaning were rarely acknowledged or appreciated by the research community. This was particularly frustrating in the face of negative or non-impressive results, when a researcher may have spent a lot of time and energy on policy retrieval and analysis, only to see their paper not accepted because the findings were deemed to be not the most novel or interesting. This is not a problem unique to privacy policy research;

⁷The Internet Archive is a non-profit organization dedicated to preserving and archiving Internet websites. <https://archive.org>

'behind-the-scenes' work is rarely appreciated in research publications. However, in privacy policy research there can be studies where a large part of the work is 'behind-the-scenes' effort spent on collecting and finding these privacy policies which often does not get the recognition it deserves. We hypothesize it may be due to the assumed ease of finding these policies since finding any one privacy policy might seem like a straightforward process (simply go to their website and download it); however, as our participants mentioned, when retrieving large quantities of privacy policies (e.g., hundreds, thousands, or more), or finding historical versions of privacy policies, it is anything but simple.

4.3 Policy Analysis

After locating and retrieving privacy policies, the next step is the actual analysis of the privacy policies. Our participants came from a diverse group of backgrounds and approached privacy policy analysis in vastly different ways. However, all participants shared the need to somehow go through and extract relevant information from the retrieved privacy policies. We identified three main approaches for policy analysis: some participants manually read and annotated each individual privacy policy in their sample. Others involved crowd workers in the analysis of privacy policy content; this was often done to help label large datasets of privacy policies. Lastly, some participants used or created software tools to automatically extract relevant information from or about privacy policies. Automated approaches ranged from metadata analysis or automatically calculating readability scores of privacy policies, to using NLP toolkits to process privacy policy text.

Each methodological approach had its unique challenges. However, there was one unifying theme: privacy policies are complicated, jargon-rich, lengthy documents that made analyzing them challenging and time-consuming. This presented unique research challenges as well as exacerbated problems usually associated with research and document analysis. We first discuss challenges related to the interpretation and lack of specificity of privacy policies, followed by discussing challenges specific to manual, crowdsourced, and automated policy analysis.

4.3.1 Policies Difficult to Read Through and Interpret. Perhaps the single biggest challenge of studying privacy policies that our participants brought up is that privacy policies are notoriously difficult to interpret. Almost all participants spoke to how confusing, technical, and full of jargon privacy policies can be.

In particular, the legal language and jargon in which privacy policies are written often require legal expertise to properly parse through, as P1 expanded on: *"As with any legal document, right, privacy policies are open to interpretation, and so the major challenge is, simply, the interpretation of the legal language. So, that's the fundamental limitation, and it plays out practically."* This is further complicated by how privacy policies can vary depending on the business sector; participants felt experience with one type of privacy policy may not translate to expertise with other types of privacy policies: *"So a hospital's privacy policy probably looks really different from a social media company's privacy policy, which looks very different from some amateur game developer's privacy policy"* (P4).

Participants further shared that the vagueness of terms used and how open to interpretation privacy policies are means that different people may read the same policy and come away with different interpretations of that policy. For instance, P16 voiced frustration over an experience where they and their colleagues tried analyzing a set of privacy policies: *"the conclusion from that exercise was if five people, all of which have PhDs or are working towards a PhD cannot agree on an interpretation what hope do you have for lay people"* concluding that *"nobody really knows what the right interpretation is."* This interpretability issue extended to people with legal expertise, as has also been shown in prior studies on experts' interpretations of privacy policies [36]. P22 shared that *"we had hired law students who were English native speakers [...] even those law students actually had at some point problem with interpreting what actually, let's say in a long sentence, the data practices, what they are actually about, and what they are correctly referring to [...] even the leader of the project, who was a professor himself, struggled with understanding what this paragraph is trying to say."*

Some participants expressed frustration at this state of affairs, blaming companies for deliberately creating privacy policies that are hard to read and interpret. Participants argued that there are few incentives for companies to make privacy policies easily interpretable. *"The nature of privacy policies... because they are written to mostly protect the owners, the providers of the policy or service, from legal problems. They are written in a language that is hard to read, it's very long, and it doesn't have nice categorization inside based on the topics... It's hard to read for a person and it's hard to be analyzed by a machine"* (P2). P12 went further, claiming that *"privacy policies are to some extent written adversarially [...] the goal, right, to some extent is to obscure what's really happening."*

4.3.2 Lack of Specificity. Some participants spoke to the difficulties of finding privacy policies or data practice descriptions that were specific to a given app, website, or service. Apps or websites sometimes link to their organization's privacy policy, which are often quite generic as they apply to all products of that organization. Although, in some cases, such policies might contain links to more product-specific privacy policies or disclosures. For instance, P14 shared: *"We found [privacy policies] were generally quite hard to find in at least like a quarter of them and then that either meant going to the developer's website or a parent company website or stumbling upon it somewhere completely different, or when you found a privacy [policy] it was very general and not necessarily specific to the function of that application."* Participants noted that these generic privacy policies that applied to a range of products made it difficult to find information about data practices specific to a given product—an issue for both researchers and consumers attempting to understand a given product's data practices and privacy implications.

This lack of specificity is further exacerbated when there is a single umbrella privacy policy that does not just apply to all of an organization's products, but rather to all of a parent organization's business divisions across multiple legal jurisdictions. Three participants (P10, P11, P26) mentioned encountering this challenge while studying websites or mobile apps that are owned by large corporate entities, such as Disney and Tencent. P26 mentions that these types of policies *"muddles it to the point where it's like, well, how does this apply to your specific website, to your specific standards, because*

what [company] does for [company service] is different than what [another service] will do where it's directly interactive." P14 expanded on these frustrations: "There wasn't a perfect correspondence between app and privacy policy because sometimes the privacy policy existed for the app's parent company and would apply to the multiple app or website offerings. Sometimes the policy was, so I think in the best case, it was attached to the app in the app store and specific to the app. Sometimes, though, it was hosted on the developer's website and applied to all that developer's products and sometimes it was even a higher level, at some parent company, and so it kind of, in some sense counted as having a privacy policy but was not always as specific or relevant to that product as we thought it should be."

Organizations using generic or umbrella policies for all their products also led to our participants expressing concerns about representation and skew in collected datasets. For example, when studying top-ranked websites or apps, multiple of them may be offered by the same (parent) organization and thus reference the same privacy policy. P8 mentioned having to manually re-check their collected dataset to remove duplicates because of this. At the same time, removing duplicates comes at the risk of under-representing data practices of apps/websites in the dataset.

4.3.3 Distrust in Policies Reflecting Actual Data Practices. Multiple participants expressed that they did not trust that privacy policies accurately reflected a company's or product's data practices, with some participants doubting the value of privacy policies as compared to code analysis. P11 explained: "from my point of view as a privacy or security researcher, I can only trust the behaviors described or discovered in the byte code." Similarly, P17 expressed interest in expanding their area of research to verifying whether apps complied with privacy policies; in their eyes, privacy policies offered "promises" about what a company did with users' data, but not necessarily guarantees.

4.3.4 Time-Consuming for Researchers to Manually Analyze Privacy Policies. In addition to the common challenges regarding policy interpretability discussed above, participants discussed multiple challenges specific to their methodological choice for privacy policy analysis. We first discuss challenges in the manual analysis of privacy policies, followed by challenges in crowdsourcing and automating privacy policy analysis in Sections 4.3.5 and 4.3.6.

Participants who engaged in manual policy analysis followed different approaches. Some participants took a grounded theory approach, identifying and developing relevant concepts as they read through privacy policies. Others started with an existing framework or codebook that they either developed themselves or adopted from the literature and read privacy policies with the goal of coding/annotating them. While the second approach allowed for more focused reading, for both approaches participants noted the amount of time required for reading through and annotating privacy policies manually. As already mentioned in Section 4.1.3, this required participants to severely limit the scale of their analysis. Even participants with legal expertise, who presumably are well-versed with legal language, mentioned needing to scale down the number of privacy policies they could use in their study.

Furthermore, the interpretability and vagueness of privacy policies meant that even when one read through a privacy policy, it was unclear whether the results were reliable or accurate. Some

participants addressed this concern by having multiple researchers independently annotate the same privacy policy and then come together to resolve differences and come to a shared interpretation. This approach added further time demands and also involved training others in consistently using a defined codebook. While challenges of interpretation and inter-rater reliability approaches are common aspects of qualitative analysis [28], participants noted how these difficulties were exacerbated by the complicated, vague, and jargon-rich nature of privacy policies.

One option participants mentioned for scaling up their efforts was hiring outside help to analyze or annotate privacy policies, such as hiring student workers or legal experts. Hiring students for policy analysis required additional time for training and flexibility. Furthermore, participants who worked with students felt it was necessary to have multiple students annotate policies for reliability reasons: "So we allowed one particular privacy policy to be labeled by multiple students so because for one student maybe their labeling would be error prone so because even though all of them were trained, the same way it could still be labeled or interpreted differently" (P5). Multiple participants considered hiring outside legal experts to analyze privacy policies. Participants mostly agreed that legal expertise would provide the highest quality of analysis; however, this option was considered prohibitively expensive by many and was therefore rarely pursued.

As a result of these challenges and constraints, manual analysis studies were often smaller in scale, e.g., involving tens of privacy policies. By focusing on a smaller set of policies, participants felt they could provide rich, qualitative insights as the researchers would have actually read through the privacy policies themselves. However, for research questions requiring analysis of larger amounts of privacy policies, participants considered manual analysis alone infeasible. While this is true for many qualitative studies, the unique complexity and difficulty of reading through privacy policy documents make manually analyzing privacy policies particularly time-consuming and so extremely limiting.

4.3.5 Crowdworkers Unreliable and/or Expensive. Some participants reported leveraging crowdsourcing as a cost-effective approach to scaling up privacy policy analysis, e.g., by posting annotation tasks on crowdsourcing platforms such as Amazon Mechanical Turk. In crowdsourcing policy annotations, the main challenge again was the unique difficulty and vagueness of privacy policies: participants felt that crowd workers either would need extensive training to produce useful annotations given the complexity of the privacy policies or the researchers needed to expend extensive time and effort on designing crowdsourcing tasks that would produce reliable results. For instance, P16 said it was possible to use crowd workers as long as one segmented the policies in an appropriate manner: "The key thing was if you give the crowd workers big tasks you'll get a lot of noise right, but if you try to break them into like into really small chunks, which are very useful, then it should not be an issue."

Another approach was to give the same task to multiple crowd workers to increase reliability. Often the number of crowd workers would be higher compared to the number of annotators involved in the manual approaches discussed above. However, participants discussed the need for finding meaningful tradeoffs between data

reliability and crowdsourcing costs, as P21 explains: *"In terms of crowd-sourced workers because the quality of the annotation is also one factor in like we have ten workers, and they normally distributes all over the place, then the noise is too much. Yeah, of course, we can go back to the cost like we can hire 10,000 workers then maybe that would make a big difference. But this means we spend 10,000 dollars every time we run, and if we did like four times it becomes 40,000 dollars."*

4.3.6 Machine Learning Approaches are Imperfect. Manually reading through privacy policies was time-consuming for participants; hiring crowdworkers was often considered unreliable and could become expensive. Therefore, multiple participants used machine learning (ML) and natural language processing (NLP) approaches to partially or fully automate privacy policy analysis by automatically identifying and extracting certain information from privacy policies. Despite the promise machine learning techniques offer in terms of automating and simplifying the analysis of privacy policies [33, 50], many participants noted difficulties in applying ML/NLP techniques and existing tools in the analysis of privacy policies.

ML/NLP techniques face technical challenges and limitations when analyzing privacy policies given the language and vagueness of privacy policy documents. Privacy policies are not machine-readable documents, and this leads to problems with inconsistent, inaccurate results. P16 stated the problems they encountered when trying to apply NLP tools to privacy policies: *"[the tools] were like really good if you had like small tangible statements it gave you like really good results, once you tried working with like privacy policies, a lot of times there was a lot of noise [...] that was because a lot of these policies were like so vague they had, like, so many different clauses, they had so many commas [...] it was not the technology breaking, it was just that these policies were written in a way that [the] intention of the policy was to be as vague as they could and just to like have you know legal upper hand in you know, in terms of user privacy."* In particular, P16 complained about the lack of accuracy of NLP tools when applied to privacy policies: *"for privacy policies they would just break so many times, or like give you an incorrect results, because as I said, even as a human you're not able to process that statement and how do you really expect a tool to you know do it accurately?"*

P11 also spoke to the difficulties of evaluating ML approaches: *"When I read other papers related to analyzing the privacy policies of mobile apps, I found that this task is very challenging, almost all of the researchers, including for our system, we can only select, for example, 20 or 30 or 50 mobile apps and then we have to manually check them and verify the false positives and negatives... due to the time limit we can only check part of the data set."*

Despite these challenges, participants still expressed hope that machine learning and NLP approaches could make privacy policy analysis easier. When asked what tools or resources they desired to make studying privacy policies easier, many participants mentioned leveraging ML/NLP tools for analysis. For example, one participant suggested a tool that converts the legal jargon of privacy policies into a more human-readable format. P14 echoed this sentiment: *"If there's some way you could like automate the lawyer part... if you could have like a machine-learned lawyer or something that would pull out these common standard phrases and almost have like*

a data dictionary where a lawyer translate that into like consumer understanding." Similarly, P6 wished for a tool to summarize privacy policies: *"Well, maybe if there was a software where you can copy privacy policies and using some AI you can extract the data that you're looking for, without having to read it all."*

4.4 Overarching Challenges

We conclude our findings by discussing overarching challenges for privacy policy research. These challenges are not specific to any single step of analyzing privacy policies, but instead, are challenges that researchers face throughout the entire research process.

4.4.1 Availability and Maintenance of Research Tools. When it comes to sharing or using research tools, particularly software, one challenging aspect mentioned by participants was the need for constant maintenance. The complicated and evolving nature of privacy policies and technology, paired with new and changing privacy regulations, meant that tools released in conjunction with research publications are often soon out-of-date unless actively maintained. As a consequence, it meant that researchers who develop and release tools to help analyze privacy policies have to spend effort not only on creating the tool in the first place but also on keeping it updated. P4 described this issue succinctly: *"So, we can get [the tool] to the finish line, but then the finish line moves."* This meant that existing research tools were often not properly maintained, meaning tools were buggy, unreliable, or unusable as a result.

Another challenge when sharing tools or datasets is that their original use case is often highly specific, as researchers tend to create them to answer a specific research question or to support a particular analysis style. Therefore, tools might not be useful and difficult to adapt for researchers who approach privacy policy analysis differently. On talking about the usefulness of annotated privacy policies, P4 mentioned *"It's difficult to imagine one data set or one annotation approach satisfying everyone's questions because there are so many things you could look for."*

4.4.2 Lack of Support for Analyzing Non-English Policies. Participants further discussed the challenges of studying privacy policies in languages other than English. First, there was a lack of tools and resources for studying policies outside of an English context. For example, most readability calculators are aimed at policies written in English. Policies written in other languages have different syntax or even characters (e.g., Chinese, Japanese, Korean) which may make English text analysis tools not applicable. Likewise, there is a lack of research and support as well as a community for studying non-English policies. P22 pointed out how relatively few papers studying non-English policies have been published: *"when you have a look at top conferences, I think there are still a lot of room for acceptance of studies and papers that focus on privacy policies other than English. There is a lot of them if you look at top conferences, still, most work that is being done is only focused on privacy policies that are in English. Yeah, I see that as a big problem personally, it may also be because of the existence of the GDPR, if there is a GDPR we need to analyze prior policies of [...] countries that are affected by GDPR, not only in English."* P22 further calls out reviewers at these conferences, claiming that on submitting a paper *"we observed that*

there are some obstacles and that reviewers are a little bit hesitant on that part [to publish multilingual analysis of privacy policies].”

P11 further talked about a lack of community around this topic. When asked about the usefulness of special interest groups as a way to collaborate with other researchers, P11 highlighted how most of these groups were Western and English-centric: *“I think, it is a little challenge to find such group and things, we know that in different countries, there are different groups that focus on privacy policy but, from my point of view, actually I found that most of them were located in America and some of them were located in Europe, but I found that few researchers were located in China or Japanese or some Asian countries, so if you have such group, I think, only American or European, such were joining this group so most Asian researchers, they cannot find such groups.”*

Lastly, researchers sometimes have to contend with country-specific challenges. One region participants mentioned as being somewhat difficult to study was China. P10 mentioned that it was more difficult to study the privacy policies of mobile apps in China since one needed to tie an ID to a SIM card purchase: *“you have to have your national ID to buy a mobile phone number [in China]. So, we don’t have that access to that SIM and we don’t have institutionally provided SIM that doesn’t related to our own identity to test, so [our work is] only focusing on Android phone and then Apple iOS.”*

4.4.3 Lack of Privacy Policy-Specific Publication Outlets. P12, who came from a computer science department, highlighted that one challenge was that it was difficult to find conferences or publication venues specifically on privacy policy analysis. Per P12, an effect of this is that one has to modify or adapt a study and its framing to make it suitable for publication in a prestigious venue with a broader range of topics. *“It’s not always obvious sort of where in the CS community this sort of research should sit. You know, the main security conferences, usually have privacy mentioned in their calls for papers but I think those are sort of written with the idea that that’s kind of technical privacy research and it’s not always apparent that you know study the policies really kind of falls into that category. And they’re still relatively few purely privacy venues that you know received the same amount of attention that you can imagine a graduate student want to publish in order to you know you put something on their CV for graduation, and you know it’s kind of the same thing from the HCI community, the big chair conferences, they often have you know privacy mentioned but, again, they also often expecting some user study involved in that”* (P12).

This issue aligns with the perception that efforts and investments in solving policy retrieval challenges often remain unappreciated and are not considered novel contributions on their own (see Section 4.2.5), despite the necessity to solve them in order to advance and scale privacy policy analysis and research.

4.4.4 Establishing Interdisciplinary Collaboration. Another overarching challenge participants brought up is the need for and difficulties in establishing interdisciplinary collaborations. For example, participants from CS departments mentioned difficulty finding collaborators outside of their fields, particularly in the legal field.

Surprisingly, the reverse was also true—participants in legal departments found it difficult to find collaborators with CS expertise. For example, P17, a legal scholar, wanted to collaborate with CS researchers to study the data flows of technologies and compare them

to what privacy policies are saying. However, they found difficulties doing so: *“I have training in law, I have training in public health, the way I view the world will always be shaped by those disciplines, and there’s so much that I don’t even know that I don’t know because I don’t have the right members on my team.”* Per P17, the difficulty in finding collaborators was a lack of interdisciplinary expertise and shared knowledge: *“I don’t even really speak the language I need to find the people I need just yet.”*

The fact that participants from different disciplines want to collaborate together but are unable to suggests that there is a disconnect between these disciplines, and bridging this gap could lead to more fruitful collaborations.

5 DISCUSSION

Our findings reveal numerous challenges researchers face when analyzing privacy policies. Based on these challenges, we lay out an agenda for improving privacy policy research. We discuss opportunities in four key areas: opportunities for research that will help mature the field; opportunities for research tool development; opportunities for companies and policymakers to better facilitate access to privacy policies for both researchers and consumers; and opportunities for building a more robust privacy policy research community that facilitates cross-disciplinary collaboration and recognizes the technical challenges involved in privacy policy research.

Pursuing our suggestions would enable higher-quality research on privacy policies and reduce barriers to conducting such research. This will not only benefit researchers carrying out this work but will also have positive impacts on the general public and society at large. Privacy policies are one of the main ways in which companies communicate their data practices to consumers and the public. The outcomes of privacy policy research can help legislators and policymakers understand company practices and develop relevant legislation to address problematic practices and help users exercise their privacy rights (for example, by developing tools that can summarize and simplify privacy policies).

5.1 Opportunities for Maturing the Field

We first discuss research opportunities that would facilitate and improve privacy policy research. We start discussing the need to understand and establish best practices for privacy policy research, followed by opportunities for studying previously-understudied privacy policies.

5.1.1 Establishing Best Practices. One of the original goals of our study was to understand how researchers analyze privacy policies and identify respective best practices. To our surprise, we discovered there were *no* clearly established best practices for studying privacy policies. Few guidelines or methodological frameworks exist for researchers to inform their approaches—instead, our participants mostly tried figuring out on their own what the best approach might be and how to overcome challenges, hence the great variation in terms of how participants approached privacy policy analysis.

This lack of established best practices can severely hinder research progress for individual researchers as well as slow down advancements in the field. A lot of effort is spent on trying to determine what is the right approach to use, and inevitably there is

duplication of work and wasted effort, such as by independently building tools to find and analyze privacy policies.

This suggests an opportunity for the research community to develop and establish best practices and workflows for privacy policy research, including ways to overcome some of the challenges identified here. Our findings demonstrate a need for increased exchange among researchers on what approaches have worked or failed, for standardizing practices and benchmarks, as well as for communal efforts to work together on solving identified policy retrieval and analysis challenges.

5.1.2 Studying the Long Tail of Privacy Policies. Many of our participants chose to study the privacy policies of widely used (or highly ranked) websites, apps, and services. This makes sense: more popular apps are used by more people, so findings concerning popular apps and companies are more impactful. However, this raises the question of whether, as a community, we are potentially understudying privacy policies and data practices of less popular offerings. As our findings show, this problem partially stems from scalability tradeoffs: if only a limited number of policies can be studied, it makes sense to choose more popular services. If these scalability issues can be solved by making large-scale analysis easier and less expensive through the reuse of resources, it becomes possible to more holistically and consistently study the long tail of privacy policies (e.g., how do small companies without in-house legal departments handle user data?). In a similar vein, we should create and value alternative metrics and filtering criteria for popularity. The discontinuation of Alexa, a popular tool for website popularity rankings, makes finding alternative ways of selecting privacy policies more critical [21].

5.1.3 Fostering Analysis of Non-English Policies. Our participants mentioned the lack of research studying non-English privacy policies, as well as the heightened difficulties in studying non-English privacy policies. This points to a need to enable and conduct more multi-lingual and non-English privacy policy research. Across the globe, a significant portion of user data is governed by companies in non-English speaking countries, as well as with policies written in non-Latin alphabets (e.g., China, South Korea, Russia, etc.). By excluding non-English privacy policies from analysis, we get an incomplete picture of the global privacy policy landscape.

5.1.4 Creating Datasets of Historical Privacy Policies. Privacy policy datasets were popular among participants, given the time and energy they would save in finding policies and labeling data (for datasets of annotated policies). One type of dataset participants wanted, but could not find, was datasets of historical privacy policies. Such historical datasets would be useful for tracking changes in privacy policies over time or measuring the impact of legislation. There is already some work in this area, such as PrivaSeer [44], the Princeton-Leuven Longitudinal Corpus [3], or TOSBack;⁸ however, these resources are not yet extensive and do not provide complete historical records of privacy policies' versions. The lack of a go-to resource for privacy policies (except the WayBack Machine) suggests a real opportunity for a repository of such policies.

⁸<https://tosback.org/>

5.2 Opportunities for Research Tools

We next discuss opportunities for research tool development. We start by discussing types of tools that would benefit the community as well as opportunities for increasing their adoption. We conclude with ways to better incentivize tool maintenance.

5.2.1 Holistic Policy Retrieval Tools. One area participants struggled with was finding and retrieving privacy policies, in particular at a large scale. Many participants employed automated crawlers; however, these crawlers ran into several issues, notably identifying which pages are privacy policies, given a lack of standardized privacy policy locations. Similarly challenging was navigating differently formatted privacy policies and extracting policy text from various document formats.

Rather than researcher teams building their own crawlers and toolchains from scratch, a community effort could substantially reduce duplication of work and yield more robust tools, such as through an open-source crawler. This crawler would have to overcome several challenges, including being able to navigate through a website's domain name and recognize when it has reached a privacy policy page; identify the format the policy is written in (e.g., PDF or HTML; whether it is dynamically loaded by JavaScript), and convert the text into a single format; and be able to identify links to external privacy policies and extract their text. For mobile apps, the tool would need to not only crawl through an application store and find privacy policy links; it would also need to navigate to those links to identify which ones are broken, or which ones lead to a company's website and not specifically the privacy policy. Moreover, it could allow researchers to specify selection criteria (e.g., number of downloads, app categories, keywords). There have been some efforts to create these processes, such as the work by Hosseini et al., who created a toolchain to collect privacy policies and prepare them for research [20]. We applaud these efforts and encourage more work in this area.

5.2.2 Tools for Policy-Code Consistency Analysis. Similarly, there is a need and opportunity for tools that enable comparison of data practices stated in a privacy policy with an app or service's actual practices. While substantial research exists on analyzing the consistency of policy statements with an app or service behavior (e.g., [10, 42, 46, 55]) most projects built their infrastructures from scratch, with little to no reuse across research groups. Furthermore, respective technological approaches often remain out of reach for non-CS researchers analyzing privacy policies, limiting the impact of those technical contributions. These tools could also serve policymakers, regulators, and consumers, by allowing these stakeholder groups to audit and see to what extent companies' policies are an accurate disclosure of their data practices.

5.2.3 Addressing Barriers to Adoption of Machine Learning Tools. Whilst creating machine learning approaches and tools to automate analysis of privacy policies is important, an arguably more critical step is addressing barriers that prevent researchers from using these tools. The lack of widespread adoption of automated privacy policy tools beyond the researchers developing them suggests that there are many challenges in this area. Advancing the performance and accuracy of ML-based analysis approaches might not be enough

if barriers to their adoption persist. Based on our interviews, we identify two main challenges.

The first barrier is the accuracy of ML-based approaches and tools, or more specifically, *knowing* the accuracy of these tools. Participants struggled with assessing the accuracy and reliability of various existing NLP toolkits. This could be particularly challenging for non-technical users who may struggle to understand the technical explanations for their usefulness. Tool creators should release comprehensive documentation, detailing their tools' benefits and limitations. Creators should not only mention standard evaluation metrics such as precision and recall but should also explain what these metrics mean in the context of their tool in plain language. Non-technical researchers would benefit greatly from examples of the limitations and the edge cases of a tool's performance, as well as examples of when it works well. There is further an opportunity here for developing benchmarks for privacy policy analysis tools.

The second barrier is *usability* of these tools. If made available, machine learning tools are often provided as GitHub repositories one must clone, download, and then execute or software packages one must import into a computer program. At times, this may require modifying the software if the software package is out of date. While this process may be feasible (albeit often still time-consuming) for researchers with computer science backgrounds, our findings indicate that non-technical researchers find it hard to impossible to use these tools. There are tools that are presented in more usable formats; for example, PrivacyCheck and Opt-Out Easy exist as browser extensions.⁹ However, these tools are often aimed at individual consumers, and may not be built for researchers looking to analyze a large set of privacy policies. Understanding usability pain points of researchers and how to increase the usefulness of tools for researchers would facilitate their adoption.

5.2.4 Appraising Large Language Models. The popularization of chatbots based on large language models (LLM) such as ChatGPT¹⁰ raises the question as to whether researchers can or will turn to them to help with the analysis of privacy policies. There may be a temptation, particularly for non-technical researchers, to use these tools, which may produce policy interpretations of unknown accuracy but are far more usable and intuitive than existing ML/NLP tools for policy analysis. Thus, research is needed to appraise LLMs and AI chatbots to understand their accuracy and usefulness for privacy policy analysis, as well as leverage and advance LLM approaches for privacy policy analysis. Furthermore, researchers should take a lesson from the ease of use that AI chatbots offer and create research tools that are as easy to use as an AI chatbot. Otherwise inexperienced or non-technical researchers will inevitably turn to potentially ill-suited off-the-shelf AI chatbots for help.

5.2.5 Incentivizing Research Tool Maintenance. Participants were often frustrated with poorly maintained research tools, which often led them to build their own toolchains instead of reusing existing ones. For those considering developing tools for the wider privacy

research community, we argue that devoting resources to the development of a maintenance plan should be considered a critical step in development, to avoid the risk of immediate obsolescence.

However, this maintenance can be challenging, not just because of the complexity of the task, but because of incentive structures for researchers. Currently, there is little recognition of the value of tool maintenance; time spent maintaining tools is time that could be spent writing new papers or applying for grants, which can offer more career opportunities.

As such, there is a need to create incentive structures that reward tool maintenance beyond current levels. For example, funding agencies could provide more grant programs specifically designated for research tool development and maintenance. Universities and departments with resources could create infrastructure to help researchers better maintain and develop existing useful tools (e.g., funds to hire full-time developers to maintain projects at competitive salaries), as well as recognize such efforts as well as their impact on the research community in tenure and promotion.

5.3 Opportunities for Industry and Policy

Here we discuss opportunities for companies and policymakers to improve the landscape for privacy policy research, including small but effective standardization steps. Importantly, realizing these opportunities would not just benefit researchers but also increase transparency for consumers.

5.3.1 Standardizing Privacy Policy Location and Metadata. Our participants called for increased privacy policy standardization, including standardizing privacy policies' structure, terminology, format, and location. If companies standardized their privacy policies, analyzing them would be vastly simpler, as researchers could focus on a single workflow for retrieving privacy policies, instead of having to account for and adapt to unending edge cases and individualized documents. Moreover, standardization would facilitate the use of tools to automate privacy policy analysis and make it easier for consumers and regulators to evaluate companies' privacy policies.

Unfortunately, standardizing privacy policies is difficult to implement in practice [13]. Calls for standardizing privacy policies stretch back to the early 2000s [5]. Various researchers have attempted to facilitate this standardization by designating standards and approaches for writing privacy policies, but none have yet achieved universal or widespread adoption (e.g., see [11, 12, 19]). The difficulty of translating privacy policies into a standardized, machine-readable language—with potential legal implications for the company—and the lack of incentives for companies to adopt a unified standard likely explains why such measures have not been adopted [9].

While standardizing all aspects of a privacy policy will be challenging, this does not imply that pursuing standardization is purposeless or quixotic. There are smaller, yet meaningful, steps that can be taken to reel in the chaotic diversity in privacy policies' format and contents. One example, inspired by robots.txt and sitemaps, would be for websites to post a *privacy.txt* document that includes basic information about their privacy policy. This approach has also been proposed by Utz et al. [47] to facilitate notifying companies of privacy issues. Such a *privacy.txt* document would be published in a website's root directory (e.g., <https://www.website.com/privacy.txt>)

⁹PrivacyCheck extension: <https://chrome.google.com/webstore/detail/privacycheck/poobeppenopkcbjejfjenbiepifcbclg>; Opt-out Easy extension: <https://chrome.google.com/webstore/detail/opt-out-easy/hikefgklfabieecchanbafeficjojik>

¹⁰<https://openai.com/blog/chatgpt>

and could contain, at a bare minimum, the URL to the website's privacy policy; the format of the privacy policy (e.g., HTML, PDF); the language of the privacy policy; and when the policy was last updated. Additional privacy-related information, such as required opt-outs, links to data access portals or requests, or privacy/DPO contact information could be easily added.

This form of standardization is relatively simple to implement since it only requires listing already-available information without interfering with the writing of the actual policy documents. Yet, widespread adoption of this approach would greatly speed up privacy policy analysis; for instance, by simplifying the automatic crawling of privacy policies from websites. Rather than having to manually or heuristically find the privacy policy, one could simply access a website's `privacy.txt` and retrieve the policy from its listed URL, together with relevant metadata.

5.3.2 Plaintext Versions of Privacy Policies. Our findings revealed a potential tension between attempts to make policies more readable and approachable for consumers, e.g., through dynamic content, and access to the full privacy policy for researchers and regulators. Dynamic and interactive policies that use JavaScript to reveal details 'as needed' and directed by the user may make the policies more transparent and usable for consumers [37]. However, this makes downloading and auditing privacy policies hard; for researchers, a structured plaintext version of the policy would be more useful. A solution could be for websites to employ *progressive enhancement*; that is, ensuring that all privacy policy text is accessible without needing to have JavaScript enabled.¹¹ This would allow researchers and regulators to easily collect and access the contents of a privacy policy, while still allowing for an inviting and dynamic version of the privacy policy. This has the added benefit of increasing accessibility for users who do not have JavaScript enabled.

5.3.3 Historical Versions of Privacy Policies. Finding historical versions of privacy policies is challenging. This not only makes it difficult for researchers to analyze these policies but also for consumers who may want to know what data practices they agreed to in the past. Alongside researchers creating historical datasets of privacy policies (see Section 5.1.4), companies should make all prior versions of their privacy policies readily available, and privacy legislation, as part of transparency requirements, could require that companies maintain public archives of their privacy policy versions. Links to these prior versions could also be published in the `privacy.txt` document mentioned in Section 5.3.1.

5.4 Building Community

Finally, we discuss opportunities for fostering an interdisciplinary research community around privacy policy analysis.

5.4.1 Privacy Policy-Specific Venues. A key takeaway from our work is the need for a more robust and integrated privacy policy analysis community. One solution to this could be the creation of conferences or journals with a dedicated focus on privacy policy analysis. Currently, privacy policy research *can* be published in top-tier venues, but as noted by P12, these venues are not focused

on privacy policies but rather serve an existing research community (e.g., privacy technologies, security, HCI, law). This leads to disciplinary silos with little interaction between researchers with technical, legal, public policy, and social science backgrounds. A dedicated venue could bring researchers from different disciplines together, facilitating collaboration and interdisciplinary learning regarding gained insights, research questions, and methodological approaches and advancements. Moreover, a dedicated venue could explicitly value and recognize research infrastructure contributions that may be overlooked or dismissed in more general venues. A community focused on studying privacy policies would more likely recognize the challenges in effectively collecting and analyzing privacy policies, as well as value respective contributions.

5.4.2 Fostering Interdisciplinary Collaboration. We found that our participants experienced a lack of, and a desire for, interdisciplinary collaboration. One way to overcome these issues are privacy policy-specific conferences which would bring together experts of various disciplines (see Section 5.4.1). One pitfall to avoid in pursuing such collaborations is viewing collaborators from other disciplines as mere 'grunt workers' on a project. For example, CS researchers should not view collaboration with legal scholars as 'someone to read through these privacy policies for me.' Similarly, legal scholars should not view collaboration with CS scholars as 'someone who builds a crawler or ML extraction tool for me.' When establishing cross-disciplinary collaborations, it is important to find meaningful questions that are truly interdisciplinary, i.e., that resonate with the different disciplines involved. For example, a research question focused on the interpretability of privacy policies might mutually interest both legal scholars and NLP/CS researchers.

6 CONCLUSION

We reported on a study involving semi-structured interviews with 26 privacy policy researchers from various disciplines. Our findings show that researchers experience many challenges in the selection of privacy policies, policy retrieval, and policy analysis, as well as a range of overarching challenges in the research process. Our findings further reveal that despite at least a decade's worth of research on privacy policies few best practices for privacy policy research have emerged, and re-use of research infrastructure and tools is scant, resulting in extensive duplicate effort by researchers. However, our interviews also provide insights into various opportunities for improving the status quo, both in terms of concrete research directions and structural changes. Fostering the interdisciplinary community around privacy policy research is crucial for maturing the field, as is incentivizing researchers to invest time and effort into building out and maintaining reusable and robust policy analysis tools and infrastructure.

ACKNOWLEDGMENTS

We thank our participants for their valuable time and insights. We thank the members of SPILab, Misha Teplitskiy, and Rahaf Alharbi for feedback on the study design and helping test out study materials. We further thank Martin Degeling, Anne Oeldorf-Hirsch, Norman Sadeh, Peter Story, Christine Utz, Henry Hosseini, and Kat Roemmich for their feedback and input on earlier drafts of this

¹¹https://developer.mozilla.org/en-US/docs/Glossary/Progressive_Enhancement

paper; as well as our anonymous shepherd and reviewers for their constructive feedback.

This material is based upon work supported by the National Science Foundation under Award No. 2105734, 2105736, and 2105745 (“Collaborative Research: SaTC: CORE: Medium: A Large-Scale, Longitudinal Resource to Advance Technical and Legal Understanding of Textual Privacy Information”).

REFERENCES

- [1] Wasi Ahmad, Jianfeng Chi, Yuan Tian, and Kai-Wei Chang. 2020. PolicyQA: A Reading Comprehension Dataset for Privacy Policies. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 743–749. <https://doi.org/10.18653/v1/2020.findings-emnlp.66>
- [2] Manar Alohal and Hassan Takabi. 2016. Better Privacy Indicators: A New Approach to Quantification of Privacy Policies. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*. USENIX Association, Denver, CO. <https://www.usenix.org/conference/soups2016/workshop-program/wpi/presentation/alohaly>
- [3] Ryan Amos, Gunes Acar, Eli Lucherini, Mihir Kshirsagar, Arvind Narayanan, and Jonathan Mayer. 2021. Privacy Policies over Time: Curation and Analysis of a Million-Document Dataset. In *Proceedings of the Web Conference 2021 (WWW '21)*. Association for Computing Machinery, New York, NY, USA, 2165–2176. <https://doi.org/10.1145/3442381.3450048>
- [4] Benjamin Andow, Samin Yaseer Mahmud, Wenyu Wang, Justin Whitaker, William Enck, Bradley Reaves, Kapil Singh, and Tao Xie. 2019. PolicyLint: Investigating Internal Privacy Policy Contradictions on Google Play. In *28th USENIX Security Symposium (USENIX Security 19)*. USENIX Association, Santa Clara, CA, 585–602. <https://www.usenix.org/conference/usenixsecurity19/presentation/andow>
- [5] Sheila F. Anthony. 2013. The Case for Standardization of Privacy Policy Formats. <https://www.ftc.gov/news-events/news/speeches/case-standardization-privacy-policy-formats>
- [6] Siddhant Arora, Henry Hosseini, Christine Utz, Vinayashkhar Bannihatti Kumar, Tristan Dhellemmes, Abhilasha Ravichander, Peter Story, Jasmine Mangat, Rex Chen, Martin Degeling, Thomas Norton, Thomas Hupperich, Shomir Wilson, and Norman Sadeh. 2022. A Tale of Two Regulatory Regimes: Creation and Analysis of a Bilingual Privacy Policy Corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 5460–5472. <https://aclanthology.org/2022.lrec-1.585>
- [7] Vinayashkhar Bannihatti Kumar, Roger Iyengar, Namita Nisal, Yuanyuan Feng, Hana Habib, Peter Story, Sushain Cherivirala, Margaret Hagan, Lorrie Cranor, Shomir Wilson, Florian Schaub, and Norman Sadeh. 2020. Finding a Choice in a Haystack: Automatic Extraction of Opt-Out Statements from Privacy Policy Text. In *Proceedings of The Web Conference 2020 (Taipei, Taiwan) (WWW '20)*. Association for Computing Machinery, New York, NY, USA, 1943–1954. <https://doi.org/10.1145/3366423.3380262>
- [8] Jaspreet Bhatia, Morgan C. Evans, and Travis D. Breaux. 2019. Identifying Incompleteness in Privacy Policy Goals Using Semantic Frames. *Requirements Engineering* 24, 3 (Sept. 2019), 291–313. <https://doi.org/10.1007/s00766-019-00315-y>
- [9] Reuben Binns. 2014. Standardised Privacy Policies: A Post-mortem and Promising Developments. In *W3C Privacy Workshop: Privacy and User-Centric Controls*. <https://www.w3.org/2014/privacyws/pp/Binns.pdf>
- [10] Duc Bui, Yuan Yao, Kang G. Shin, Jong-Min Choi, and Junbum Shin. 2021. Consistency Analysis of Data-Usage Purposes in Mobile Apps. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security (Virtual Event, Republic of Korea) (CCS '21)*. Association for Computing Machinery, New York, NY, USA, 2824–2843. <https://doi.org/10.1145/3460120.3484536>
- [11] João Caramujo, Alberto Rodrigues da Silva, Shaghayegh Monfared, André Ribeiro, Pável Calado, and Travis Breaux. 2019. RSL-IL4Privacy: A Domain-Specific Language for the Rigorous Specification of Privacy Policies. *Requirements Engineering* 24, 1 (March 2019), 1–26. <https://doi.org/10.1007/s00766-018-0305-2>
- [12] Lorrie Faith Cranor. 2003. P3P: Making Privacy Policies More Useful. *IEEE Security & Privacy* 1, 6 (Nov. 2003), 50–55. <https://doi.org/10.1109/MSECP.2003.1253568>
- [13] Lorrie Faith Cranor. 2012. Necessary but not sufficient: Standardized mechanisms for privacy notice and choice. *J. on Telecomm. & High Tech. L.* 10 (2012), 273.
- [14] Gitanjali Das, Cynthia Cheung, Camille Nebeker, Matthew Bietz, and Cinnamon Bloss. 2018. Privacy Policies for Apps Targeted Toward Youth: Descriptive Analysis of Readability. *JMIR mHealth and uHealth* 6, 1 (Jan. 2018), e7626. <https://doi.org/10.2196/mhealth.7626>
- [15] Martin Degeling, Christine Utz, Christopher Lentzsch, Henry Hosseini, Florian Schaub, and Thorsten Holz. 2019. We Value Your Privacy ... Now Take Some Cookies. *Informatik Spektrum* 42, 5 (Oct. 2019), 345–346. <https://doi.org/10.1007/s00287-019-01201-1>
- [16] Benjamin Fabian, Tatiana Ermakova, and Tino Lentz. 2017. Large-Scale Readability Analysis of Privacy Policies. In *Proceedings of the International Conference on Web Intelligence (WI '17)*. Association for Computing Machinery, New York, NY, USA, 18–25. <https://doi.org/10.1145/3106426.3106427>
- [17] Tao Fu. 2019. China’s Personal Information Protection in a Data-Driven Economy: A Privacy Policy Study of Alibaba, Baidu and Tencent. *Global Media and Communication* 15, 2 (Aug. 2019), 195–213. <https://doi.org/10.1177/1742766519846644>
- [18] Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G. Shin, and Karl Aberer. 2018. Polisis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning. In *27th USENIX Security Symposium (USENIX Security 18)*. USENIX Association, Baltimore, MD, 531–548. <https://www.usenix.org/conference/usenixsecurity18/presentation/harkous>
- [19] Martin Henze, Jens Hiller, Sascha Schmerling, Jan Henrik Ziegeldorf, and Klaus Wehrle. 2016. CPPL: Compact Privacy Policy Language. In *Proceedings of the 2016 ACM on Workshop on Privacy in the Electronic Society (WPES '16)*. Association for Computing Machinery, New York, NY, USA, 99–110. <https://doi.org/10.1145/2994620.2994627>
- [20] Henry Hosseini, Martin Degeling, Christine Utz, and Thomas Hupperich. 2021. Unifying Privacy Policy Detection. *Proc. Priv. Enhancing Technol.* 2021, 4 (2021), 480–499.
- [21] Kishanu Karmakar. 2021. Alexa.Com Website Ranking Service Will Be Discontinued by Amazon. <https://talktoiconic.com/alexa-com-website-ranking-service-will-be-discontinued-by-amazon/>
- [22] Moniba Keymanesh, Micha Elsner, and Srinivasan Sarthasathary. 2020. Toward Domain-Guided Controllable Summarization of Privacy Policies. In *Proceedings of the 2020 Natural Language Processing (NLLP) Workshop*. ACM, New York, NY, USA, 18–24. <http://ceur-ws.org/Vol-2645/paper3.pdf>
- [23] Logan Lebanoff and Fei Liu. 2018. Automatic Detection of Vague Words and Sentences in Privacy Policies. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 3508–3517. <https://doi.org/10.18653/v1/D18-1387>
- [24] Timothy Libert. 2018. An Automated Approach to Auditing Disclosure of Third-Party Data Collection in Website Privacy Policies. In *Proceedings of the 2018 World Wide Web Conference (WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 207–216. <https://doi.org/10.1145/3178876.3186087>
- [25] Fei Liu, Nicole Lee Fella, and Kexin Liao. 2018. Modeling Language Vagueness in Privacy Policies using Deep Neural Networks. *CoRR* abs/1805.10393 (2018). [arXiv:1805.10393](https://arxiv.org/abs/1805.10393) <http://arxiv.org/abs/1805.10393>
- [26] Florian Mathis, Kami Vaniea, and Mohamed Khamis. 2022. Prototyping Usable Privacy and Security Systems: Insights from Experts. *International Journal of Human-Computer Interaction* 38, 5 (March 2022), 468–490. <https://doi.org/10.1080/10447318.2021.1949134>
- [27] Alecia M. McDonald and Lorrie Faith Cranor. 2008. The Cost of Reading Privacy Policies 2008 Privacy Year in Review. *I/S: A Journal of Law and Policy for the Information Society* 4, 3 (2008), 543–568. <http://hdl.handle.net/1811/72839>
- [28] Matthew B Miles, A Michael Huberman, and Johnny Saldaña. 2014. *Qualitative Data Analysis: A Methods Sourcebook*. Sage publications, Thousand Oaks, CA, USA.
- [29] Jayashree Mohan, Melissa Wasserman, and Vijay Chidambaram. 2019. Analyzing GDPR Compliance Through the Lens of Privacy Policy. In *Heterogeneous Data Management, Polystores, and Analytics for Healthcare (Lecture Notes in Computer Science)*. Vijay Gadeppally, Timothy Mattson, Michael Stonebraker, Fusheng Wang, Gang Luo, Yanhui Laing, and Alevtina Dubovitskaya (Eds.). Springer International Publishing, Cham, 82–95. https://doi.org/10.1007/978-3-030-33752-0_6
- [30] Razieh Nokhbeh Zaeem, Ahmad Ahabab, Josh Bestor, Hussam H. Djadi, Sunny Kharel, Victor Lai, Nick Wang, and K. Suzanne Barber. 2022. PrivacyCheck v3: Empowering Users with Higher-Level Understanding of Privacy Policies. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining (Virtual Event, AZ, USA) (WSDM '22)*. Association for Computing Machinery, New York, NY, USA, 1593–1596. <https://doi.org/10.1145/3488560.3502184>
- [31] Razieh Nokhbeh Zaeem, Safa Anya, Alex Issa, Jake Nimergood, Isabelle Rogers, Vinay Shah, Ayush Srivastava, and K. Suzanne Barber. 2020. PrivacyCheck v2: A Tool That Recaps Privacy Policies for You. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (Virtual Event, Ireland) (CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 3441–3444. <https://doi.org/10.1145/3340531.3417469>
- [32] Anne Oeldorf-Hirsch and Jonathan A. Obar. 2019. Overwhelming, Important, Irrelevant: Terms of Service and Privacy Policy Reading among Older Adults. In *Proceedings of the 10th International Conference on Social Media and Society (SMoSociety '19)*. Association for Computing Machinery, New York, NY, USA, 166–173. <https://doi.org/10.1145/3328529.3328557>
- [33] Abhilasha Ravichander, Alan W Black, Thomas Norton, Shomir Wilson, and Norman Sadeh. 2021. Breaking Down Walls of Text: How Can NLP Benefit Consumer Privacy?. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, Vol. 1. Association for Computational Linguistics (ACL),

- Virtual, Online, 4125–4140. <https://aclanthology.org/2021.acl-long.319>
- [34] Abhilasha Ravichander, Alan W Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. 2019. Question Answering for Privacy Policies: Combining Computational and Legal Perspectives. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 4949–4959. <https://doi.org/10.48550/arXiv.1911.00841>
- [35] Ronak Razavisousan and Karuna P. Joshi. 2021. Analyzing GDPR compliance in Cloud Services' privacy policies using Textual Fuzzy Interpretive Structural Modeling (TFISM). In *2021 IEEE International Conference on Services Computing (SCC)*. IEEE, Chicago, IL, USA, 89–98. <https://doi.org/10.1109/SCC53864.2021.00021>
- [36] Joel R. Reidenberg, Travis Breaux, Lorrie Faith Cranor, Brian French, Amanda Grannis, James T. Graves, Fei Liu, Alecia McDonald, Thomas B. Norton, N. Cameron Ramanath, Rohanand Russell, Norman Sadeh, and Florian Schaub. 2015. Disagreeable Privacy Policies: Mismatches between Meaning and Users' Understanding. *Berkeley Technology Law Journal* 30 (2015), 39. <http://dx.doi.org/10.2139/ssrn.2418297>
- [37] Daniel Reinhardt, Johannes Borchard, and Jörn Hürtienne. 2021. Visual Interactive Privacy Policy: The Better Choice?. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–12. <https://doi.org/10.1145/3411764.3445465>
- [38] Lisa Rosenfeld, John Torous, and Ipsit V. Vahia. 2017. Data Security and Privacy in Apps for Dementia: An Analysis of Existing Privacy Policies. *The American Journal of Geriatric Psychiatry* 25, 8 (Aug. 2017), 873–877. <https://doi.org/10.1016/j.jagp.2017.04.009>
- [39] N. Cameron Russell, Joel R. Reidenberg, and Sumyung Moon. 2018. Privacy in Gaming. *Fordham Intell. Prop. Media & Ent.* 29, 61 (March 2018), 89 pages. <https://doi.org/10.2139/ssrn.3147068>
- [40] Norman Sadeh, Alessandro Acquisti, Travis D Breaux, Lorrie Faith Cranor, Alecia M McDonald, Joel R Reidenberg, Noah A Smith, Fei Liu, N Cameron Russell, Florian Schaub, et al. 2013. The usable privacy policy project. In *Technical report, Technical Report, CMU-ISR-13-119*. Carnegie Mellon University. <http://reports-archive.adm.cs.cmu.edu/anon/isr2013/abstracts/13-119.html>
- [41] Yan Shvartzshnaider, Noah Aphthorpe, Nick Feamster, and Helen Nissenbaum. 2019. Going against the (Appropriate) Flow: A Contextual Integrity Approach to Privacy Policy Analysis. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7 (Oct. 2019), 162–170. <https://doi.org/10.1609/hcomp.v7i1.5266>
- [42] Rocky Slavin, Xiaoyin Wang, Mitra Bokaei Hosseini, James Hester, Ram Krishnan, Jaspreet Bhatia, Travis D. Breaux, and Jianwei Niu. 2016. Toward a Framework for Detecting Privacy Policy Violations in Android Application Code. In *Proceedings of the 38th International Conference on Software Engineering (ICSE '16)*. Association for Computing Machinery, New York, NY, USA, 25–36. <https://doi.org/10.1145/2884781.2884855>
- [43] Mukund Srinath, Soundarya Nurani Sundareswara, C. Lee Giles, and Shomir Wilson. 2021. PrivaSeer: A Privacy Policy Search Engine. In *Web Engineering: 21st International Conference, ICWE 2021, Biarritz, France, May 18–21, 2021, Proceedings (Biarritz, France)*. Springer-Verlag, Berlin, Heidelberg, 286–301. https://doi.org/10.1007/978-3-030-74296-6_22
- [44] Mukund Srinath, Shomir Wilson, and C Lee Giles. 2021. Privacy at Scale: Introducing the PrivaSeer Corpus of Web Privacy Policies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 6829–6839. <https://doi.org/10.18653/v1/2021.acl-long.532>
- [45] Noriko Tomuro, Steven Lytinen, and Kurt Hornsburg. 2016. Automatic Summarization of Privacy Policies Using Ensemble Learning. In *Proceedings of the Sixth ACM Conference on Data and Application Security and Privacy (CODASPY '16)*. Association for Computing Machinery, New York, NY, USA, 133–135. <https://doi.org/10.1145/2857705.2857741>
- [46] Rahmadi Trimananda, Hieu Le, Hao Cui, Janice Tran Ho, Anastasia Shuba, and Athina Markopoulou. 2022. OVRseen: Auditing Network Traffic and Privacy Policies in Oculus VR. In *31st USENIX Security Symposium (USENIX Security 22)*. USENIX Association, Boston, MA, 3789–3806. <https://www.usenix.org/conference/usenixsecurity22/presentation/trimananda>
- [47] Christine Utz, Matthias Michels, Martin Degeling, Ninja Marnau, and Ben Stock. 2023. Comparing Large-Scale Privacy and Security Notifications. In *PETS 2023. Proceedings on Privacy Enhancing Technologies*. <https://publications.cispa-saarland.de/3918/>
- [48] Isabel Wagner. 2023. Privacy Policies across the Ages: Content of Privacy Policies 1996–2021. *ACM Transactions on Privacy and Security* 26, 3 (May 2023), 32:1–32:32. <https://doi.org/10.1145/3590152>
- [49] Shomir Wilson, Florian Schaub, Aswath Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N. Cameron Russell, Thomas B. Norton, Eduard Hovy, Joel Reidenberg, and Norman Sadeh. 2016. The Creation and Analysis of a Website Privacy Policy Corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, 1330–1340. <https://doi.org/10.18653/v1/P16-1126>
- [50] Shomir Wilson, Florian Schaub, Frederick Liu, Kanthashree Mysore Sathyendra, Daniel Smullen, Sebastian Zimmeck, Rohan Ramanath, Peter Story, Fei Liu, Norman Sadeh, and Noah A. Smith. 2018. Analyzing Privacy Policies at Scale: From Crowdsourcing to Automated Annotations. *ACM Transactions on the Web* 13, 1 (Dec. 2018), 1:1–1:29. <https://doi.org/10.1145/3230665>
- [51] Le Yu, Xiapu Luo, Jiachi Chen, Hao Zhou, Tao Zhang, Henry Chang, and Hareton K. N. Leung. 2021. PPChecker: Towards Accessing the Trustworthiness of Android Apps' Privacy Policies. *IEEE Transactions on Software Engineering* 47, 2 (Feb. 2021), 221–242. <https://doi.org/10.1109/TSE.2018.2886875>
- [52] Razieh Nokhbeh Zaeem and K. Suzanne Barber. 2020. The Effect of the GDPR on Privacy Policies: Recent Progress and Future Promise. *ACM Transactions on Management Information Systems* 12, 1 (Dec. 2020), 2:1–2:20. <https://doi.org/10.1145/3389685>
- [53] Razieh Nokhbeh Zaeem, Rachel L. German, and K. Suzanne Barber. 2018. PrivacyCheck: Automatic Summarization of Privacy Policies Using Data Mining. *ACM Trans. Internet Technol.* 18, 4, Article 53 (aug 2018), 18 pages. <https://doi.org/10.1145/3127519>
- [54] Melvyn Zhang, Aloysius Chow, and Helen Smith. 2020. COVID-19 Contact-Tracing Apps: Analysis of the Readability of Privacy Policies. *Journal of Medical Internet Research* 22, 12 (Dec. 2020), e21572. <https://doi.org/10.2196/21572>
- [55] Sebastian Zimmeck, Peter Story, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel Reidenberg, N. Russell, and Norman Sadeh. 2019. MAPS: Scaling Privacy Compliance Analysis to a Million Apps. *Proc. Priv. Enhancing Tech.* 2019 (Jan. 2019), 66. https://ir.lawnet.fordham.edu/faculty_scholarship/1040
- [56] Sebastian Zimmeck, Ziqi Wang, Lieyong Zou, Roger Iyengar, Bin Liu, Florian Schaub, Shomir Wilson, Norman Sadeh, Steven M. Bellovin, and Joel Reidenberg. 2017. Automated Analysis of Privacy Requirements for Mobile Apps. In *Proceedings 2017 Network and Distributed System Security Symposium*. Korea Society of Internet Information, Korea, Republic of. <https://doi.org/10.14722/ndss.2017.23034>

A INTERVIEW PROTOCOL

(1) Participant background information

Before we begin talking about the specifics of studying privacy policies, we want to get a better sense of you and your research background.

- First, if you could confirm the following information.
 - Based on the paper we contacted you about, you are a JOB TITLE working at DEPARTMENT / INSTITUTION in COUNTRY. Is that correct?
 - Do you have any preferred pronouns?
- How would you characterize the research you do, and how does studying or working with privacy policies factor into your work?
 - What would you say your area of research is?
 - What discipline(s) would you say you belong to?
 - How often are you involved in projects that analyze privacy policies?

(2) Paper specifics

We now want to transition towards talking about the ins and outs of studying privacy policies. Since we are going to be asking specific questions about the process, we ask that you focus on a study where you analyzed privacy policies. For example, the paper we contacted you about, (NAME OF PAPER). However, if there is another paper or project you feel is more memorable, or you would rather talk about, you can bring up that paper instead.

I'll give you some time to think about this project—let me know when you have one chosen.

- Can you give an overview of this paper / project, and what role privacy policies played into it?
 - What was the goal of the paper?

- What did you do in the paper?
- What methods did you use?
- When was the research carried out?
- What were you trying to learn from the privacy policies?

(3) Method for studying privacy policies

Thank you for sharing! Now I want to delve a bit deeper into the specifics of working with privacy policies. For this section, we ask that you refer to (NAME OF PAPER), but if there is relevant experience from other papers, feel free to bring those up as well.

- How many (and what privacy policies) did you study?
 - How did you collect these privacy policies?
 - Where did you find these privacy policies?
 - Use existing datasets? Crawling? Selecting policies manually?
 - How did you choose what privacy policies to study?
 - Did you have any inclusion or exclusion criteria?
 - Are there specific types of policies you are trying to find / important ways of categorizing policies (industry, jurisdiction, practices)?
- How did you analyze the privacy policies?
 - How did you choose to analyze them this way?
 - What made you pick this method over other methods?
 - Do you prefer manual/qualitative analysis or automated/quantitative analysis?
 - Are there any other methods you considered using that you ended up not using? Why?
 - What did you do with the results of this analysis?
- OPTIONAL [ask if the person has developed a toolkit or a new analysis framework]: You mentioned developing [an analysis tool / a framework]. Did you share this tool / script with the broader research community?
 - Why or why not?
 - How did you share it?
 - What kind of responses did you get after releasing it?

(4) Challenges

I now want to shift focus in regards to the challenges, or difficulties, that you faced when studying these privacy policies.

- What were the major challenges or difficulties you faced when studying privacy policies?
 - How did you overcome them, if at all?
 - If you could redo the study, what would you do differently to overcome these challenges?
- On a broader level, are there any barriers or challenges preventing you from studying privacy policies in more of your work?

(5) Interest/needs for research infrastructure and resources

We've talked a lot about the work you have done, the ways in which you've worked with privacy policies, and the challenges you faced when doing so.

- What type of research infrastructure, resources, or tools do you think would make this work easier? Either existing ones or ones that do not exist yet.
 - It can be anything, as wide or as narrow that you can think of. Either something to help you find privacy policies easier, something to make the analysis go faster...

- What tasks could be automated?
- Is there any existing infrastructure you use?
- Why do you use them? What are the major drawbacks?

- Thank you so much! To wrap up the interview, I want to show you a list of possible tools, infrastructure, or utility that could possibly help researchers overcome challenges when studying privacy policies. I want you to take a look at these, and let me know which ones stand out as being especially useful or especially not useful, and any general reactions to these tools

- A search engine that would allow a researcher to search for privacy policies (e.g., a 'Google for privacy policies').
- A website where researchers could upload custom programs and scripts that aid in the analysis of privacy policies, and other researchers could download these programs and use them in their own research (e.g., 'GitHub for privacy policy analysis scripts').
- A special interest group for researchers interested in studying privacy policies.
- Longitudinal data that tracks how privacy policies have changed over time.
- A corpus of privacy policies that have been annotated by privacy and legal experts to highlight useful information about the policy (e.g., what data is being collected, data storage policies of that entity, etc.).
- Datasets or repositories of privacy policies belonging to certain industry sectors.

(6) Conclusion

Thank you so much! This conversation was really insightful. Those are almost all the questions we had for you today. We have two final questions:

- First, are there any questions or topics that you expected me to ask about but I didn't?
- And second, are there any final questions or comments you want to make on the record?

B CODEBOOK

Codes		Description
Analyzing policies		How participants analyzed privacy policies.
Crowdsourcing		The participant crowdsourced work for analyzing privacy policies and describes rationale for doing so.
Manual analysis with authors		The participant analyzed privacy policies manually with a group and describes rationale for doing so.
ML Classifier		The participant used or built an ML classifier to analyze privacy policies and describes rationale for doing so.
Outside experts		The participant used outside experts to analyze privacy policies and describes rationale for doing so.
Students		The participant analyzed privacy policies with students (undergrad/graduate) and describes rationale for doing so.
Other analysis		The participant used another method to analyze privacy policies and describes rationale for doing so.
Finding policies		How participants collected privacy policies to study.
Filtering criteria	Popular Random Other criteria	The criteria participants applied used to collect privacy policies. Participant filtered policies based on popularity. Participant filtered policies based on random selection. Participant filtered policies based on other criteria.
Automatic crawler		The participant used or built an automated crawler to collect privacy policies.
Datasets		The participant used an existing dataset to collect privacy policies.
Manual selection		The participant manually collected privacy policies.
Misc. finding policies		The participant used a different method to collect privacy policies.
Number of policies		The number of privacy policies collected during the process of finding policies.
Paper motivation		The participant's goal in studying privacy policies (e.g., what the participant was searching for, and what their study contributed).
Compliance with laws		The participant studies privacy policies to see whether they comply with relevant legislation.
Data Practices		The participant studies privacy policies to examine what they collect/retain/share.
Training ML classifiers		The participant studies privacy policies to train a ML classifier or create a new machine learning tool.
Usability of privacy policies		The participant studies privacy policies to understand their usability, such as their readability, clarity, or complexity.
Other motivation		The participant studied privacy policies for another reason.
Value of privacy policies		What participants viewed as the value of studying privacy policies.
Analyzing policies challenges		What challenges the participants faced when studying privacy policies.
Difficulty assessing jargon		Participant mentions difficulties understanding or interpreting privacy policies.
Not machine readable		Privacy policies are not written in ways that can easily be read by machines.

Finding policies challenges	Difficulties participants encountered while in the process of collecting privacy policies to study.
<p>Difficult to crawl Difficult to find Format</p> <p>Parent-level policies</p>	<p>Participant mentions difficulties crawling sites/stores. Participant mentions difficulties finding privacy policies. Participant mentions difficulties arising due to the format of the policy (e.g., difficulty getting text out of policy written using JavaScript). Participant mentions that policies point to the parent or umbrella policy rather than the specific site/app/device they are studying.</p>
General challenges	Challenges participants faced that did not have to do with either analyzing or finding policies.
<p>Convenience</p> <p>Difficulty collaborating Global challenges</p> <p>Lack of expertise</p> <p>Lack of Resources No academic home Not standardized</p> <p>Peer reviews Problems with existing tools / methods</p> <p>Limits of ML tools Time consuming</p> <p>Other challenges</p>	<p>Choices of what to do made because it was 'easier' or more convenient. Includes mentioning doing stuff because of prior work/papers/related to making work easier for the participant. Difficult to find collaborators, or collaborating in general. Challenges that arise when conducting research that is not US/EU-centered. Participant remarks to challenges having to build/create/study outside of their general knowledge. E.g., lack of legal knowledge, lack of technical knowledge to build tools, lack of familiarity with the space, etc. Lack of resources (e.g., financial, processing power, man-power). Lack of community to either publish work or to network Participant mentions issues with privacy policies not being standardized. Participant mentions issues in the peer review process Participant mentions difficulties with certain methods or problems with existing tools they use during research. Participant mentions challenges encountered with ML tools. Participant mentions it takes a lot of time and effort to study privacy policies. Other challenges not captured by the above codes.</p>
Tools	
<p>Sharing tools with the wider community Things people want Tools are improving Tools need to be 'up-to-date'</p> <p>What tools did people use</p> <p>Alexa Ranking OPP-115</p> <p>Suggested tools</p> <p>Annotated datasets</p> <p>Datasets for industry sectors</p> <p>Github for privacy policies</p>	<p>Participant's reasoning for sharing/not sharing work with the wider research community. Imagined tools that participants would want to have. Tools/research has improved over time. Participant describes the need for tools to be 'up-to-date' and maintained. Names of tools that participants used when studying, finding, or analyzing policies. Only include tool's name, and awareness of other tools. Participant used the Alexa ranking. Participant used the OPP-115 corpus. Participant answers to specific tool prompts. Participant answers and reactions to annotated datasets. This includes instances the participant mentions annotated datasets outside of the prompt. Participant answers and reactions to datasets for industry sectors. This includes instances the participant mentions datasets for industry sectors outside of the prompt. Participant answers and reactions to a GitHub for privacy policies. This includes instances the participant mentions a GitHub for privacy policies outside of the prompt.</p>

Tools [cont.]		
Suggested tools [cont.]		
	Longitudinal data	Participant answers to specific tool prompts. Participant answers and reactions to longitudinal data. This includes instances the participant mentions longitudinal data outside of the prompt.
	Search engine	Participant answers and reactions to a search engine tool. This includes instances the participant mentions search engines outside of the prompt.
	Special interest groups	Participant answers and reactions to a special interest group. This includes instances the participant mentions special interest groups outside of the prompt.