Automated Large-Scale Analysis of Cookie Notice Compliance

Ahmed Bouhoula, Karel Kubicek, Amit Zac, Carlos Cotrini,
and David Basin, *ETH Zurich*

https://www.usenix.org/conference/usenixsecurity24/presentation/bouhoula

This paper is included in the Proceedings of the 33rd USENIX Security Symposium.

August 14–16, 2024 • Philadelphia, PA, USA

978-1-939133-44-1

Open access to the Proceedings of the 33rd USENIX Security Symposium is sponsored by USENIX.

Automated Large-Scale Analysis of Cookie Notice Compliance

Ahmed Bouhoula *ETH Zurich*
Karel Kubicek *ETH Zurich*
Amit Zac *ETH Zurich*
Carlos Cotrini *ETH Zurich*
David Basin *ETH Zurich*

Abstract

Privacy regulations such as the General Data Protection Regulation (*GDPR*) require websites to inform EU-based users about non-essential data collection and to request their consent to this practice. Previous studies have documented widespread violations of these regulations. However, these studies provide a limited view of the general compliance picture: they are either restricted to a subset of notice types, detect only simple violations using prescribed patterns, or analyze notices manually. Thus, they are restricted both in their scope and in their ability to analyze violations at scale.

We present the first general, automated, large-scale analysis of cookie notice compliance. Our method interacts with cookie notices, e.g., by navigating through their settings. It observes declared processing purposes and available consent options using Natural Language Processing and compares them to the actual use of cookies. By virtue of the generality and scale of our analysis, we correct for the selection bias present in previous studies focusing on specific Consent Management Platforms (*CMP*). We also provide a more general view of the overall compliance picture using a set of 97k websites popular in the EU. We report, in particular, that 65.4% of websites offering a cookie rejection option likely collect user data despite explicit negative consent.

1 Introduction

More than 90% of websites track their users and collect their personal data [44]. Their reasons range from personalizing advertisements to gaining insights into how the websites are used and their visitors' demographics.

To deter privacy-intrusive practices, privacy regulations were introduced in the European Union (*EU*), including the General Data Protection Regulation (*GDPR*) and the ePrivacy Directive. These laws mandate, in particular, that websites inform users about the explicit purposes for which their data is collected. They also require websites to have a legal basis, such as user consent, for their data collection practices. This has led to the global adoption of cookie notices, which are now unavoidable when browsing the web.

To understand how websites have reacted to and adopted the GDPR, various studies have been conducted [4, 31, 38, 41, 45, 50]. However, these studies were either not general or not at scale, as they had at least one of the following major limitations. First, **they did not interact with cookie notices**. Cookie notices take a variety of shapes and forms and often include *interactive elements* that allow users to accept or reject cookies, dismiss the notice, or navigate to *layers* containing fine-grained settings. Second, some of the studies **were conducted manually**, limiting them to too few websites to provide statistically significant observations. We particularly observe this limitation in studies focusing on deceptive design patterns (or dark patterns), which undermine legal requirements by tricking users into consent against their intentions [5, 22, 25, 41, 44, 48, 52]. Finally, some automated studies **are subject to a selection bias** because they only analyze websites employing specific Consent Management Platforms (*CMP*). This bias is prevalent because it is challenging to automate the analysis of all websites given how diverse website implementations are.

Our work[1]

Our work addresses all three limitations. We automate and generalize the analysis of websites' non-compliance with the GDPR by developing a crawler that interacts with cookie notices in a general way using machine learning. Our crawler is independent of specific CMP implementations and can navigate through the notice layers.

Generalizing and scaling up the analysis of GDPR compliance requires, in particular, the automation of three tasks: navigating cookie notices, analyzing their natural language, and attributing usage to cookies. We train three machine learning models for each of these tasks. The first model classifies elements of cookie notices that users can interact with and

[1] The source code will be made available at https://ahmedbouhoula.github.io/post/automated.html.

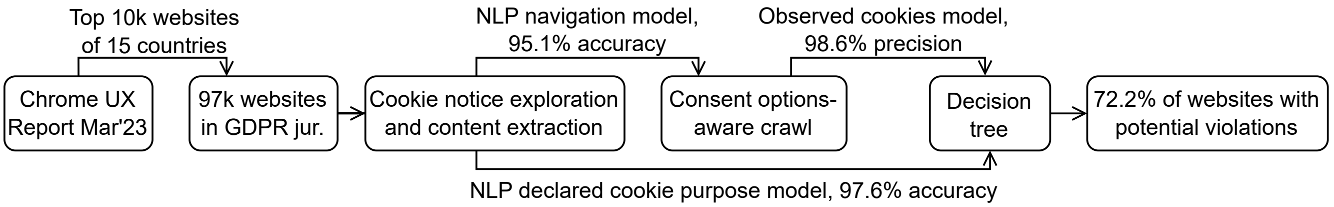

Top 10k websites of 15 countries
Chrome UX Report Mar'23
97k websites in GDPR jur.
NLP navigation model, 95.1% accuracy
Cookie notice exploration and content extraction
Consent options-aware crawl
Observed cookies model, 98.6% precision
Decision tree
72.2% of websites with potential violations
NLP declared cookie purpose model, 97.6% accuracy

Figure 1: Overview of the process involved in our study.

was trained on a dataset of 2.4k samples that we annotated. The second model detects and classifies cookie purposes in the text and was trained on a dataset provided by Santos et al. [45]. The third model extends the work of Bollinger et al. [4] and allows us to detect when a website sets cookies that require the user's consent.

Our generalization allows us to conduct a large-scale study on 97k websites in eleven EU languages. This improves upon past studies in several respects. First, our reporting is independent of the cookie notice implementation method and is therefore not prone to a selection bias like past studies. Second, by using Natural Language Processing (NLP) and heuristics, we can analyze the browser-rendered cookie notices as users observe them, unlike methods depending on the implementation of CMPs. This enables the detection of more comprehensive violations and the automated analysis of dark patterns. Third, our method is more suitable for reproducibility and long-term studies than past works depending on CMP implementations, which are subject to change over time. This will enable the analysis of compliance-shaping events. For instance, legal scholars can measure compliance before and after new regulations or court decisions, and browser vendors can observe the impact of events like the discontinuation of third-party cookies.

Using the trained models and observed usage of cookies by the websites, we report on various discrepancies associated with potential violations of users' privacy. Namely, we show that 56.7% of cookie notices do not include an option to opt out of consent, that more than 65.4% of websites with an opt-out option collect users' data despite explicit negative consent, and that 73.4% of websites do so even when users do not interact with the cookie notice. These violations were investigated by prior works with vastly different outcomes depending on the CMPs they considered. We inspect the selection bias and present results that are independent of the technologies used to implement the cookie notice. Our method also allows us to highlight a novel violation, namely that 26.1% of websites do not declare adequate data collection purposes in their cookie notices.

The generality of our approach paves the way for various types of aggregated analyses such as popularity-based and country-level analyses. As an example, we show that while more popular websites are more likely to contain cookie notices and these notices are more likely to include a "Reject" button and adequate purposes, they are also more likely to ignore the choices that users select and track them anyway.

Contributions

- We develop a general method for the automated analysis of GDPR violations and dark patterns at scale, regardless of websites' implementations.
- We conduct the largest case study to date on cookie purpose compliance, covering 97k websites. Our sample is heterogeneous and removes the selection bias specific to a subset of websites implementing certain CMPs, as is shown by our results.
- We report on potential GDPR violations and dark patterns. For instance, we find that 65.4% of websites do not respect users' negative consent and that top-ranked websites are more likely to ignore users' choices despite having seemingly more compliant cookie notices.

2 Study overview

Fig. 1 provides an overview of our study. We collected a dataset of 97k websites within the GDPR's jurisdiction using the Chrome User Experience Report (*CrUX*) [10] from March 2023. Our crawler (Section 3) visits each website and, if the detected language is one of eleven supported languages, it then searches for a cookie notice. If one is found, the crawler interacts with it, extracting its text, detecting consent options (e.g., "Reject" or "Close"), and extracting stored cookies. The detection of consent options relies on a machine learning model (Section 4.2) that classifies interactive elements—i.e., elements that users can interact with. The data collected by the crawler is further analyzed using two machine learning models: The first model (Section 4.1) predicts whether a cookie notice *declares purposes* for analytics or advertising (*AA purposes*). The second model (Section 4.3) predicts whether a website actually *uses cookies* for analytics or advertising (*AA cookies*).

The results of the crawler and machine learning models serve as input for a decision tree to uncover potential GDPR violations (Section 5.1). For example, we determine whether a website uses AA cookies despite explicitly rejecting consent and whether a website declares adequate purposes when AA cookies are used. A similar decision tree (Section 5.2) is used to uncover potential dark patterns that may deceive users and nudge them towards accepting consent.

We repeat results from previous studies that limited the analysis to websites using specific CMPs and show that they suffer from a selection bias (Section 6). We also manually inspect 500 websites and present an end-to-end evaluation of our detection of violations and dark patterns (Section 7).

3 Crawler

In this section, we describe our implementation of a web crawler that detects cookie notices, extracts their text and interactive elements, and stores the cookies observed while using the website. The crawler is implemented as an extension of the OpenWPM framework [13].

3.1 Language support

Measuring GDPR cookie consent violations requires multilingual support for websites in different EU countries and the UK.[2] Our pipeline relies on spaCy models [27] for part-of-speech tagging and on other models that we train on English datasets. For non-English texts, we use the open-source machine translation API LibreTranslate[3].

Both spaCy and LibreTranslate support the following EU languages: Danish, German, English, Spanish, Finnish, French, Italian, Dutch, Polish, Portuguese, and Swedish. Although Greek is also supported by both, LibreTranslate fails to translate basic cookie notice phrases, so we did not include Greek websites in our analysis. We summarize the language support in the Appendix in Table 3.

As the crawl starts, we detect the website's language using the langdetect[4] library, which supports all of the selected languages. We stop crawling a website when its content is not in one of the eleven selected languages.

3.2 Cookie notice detection

We first identify a set of potential HTML elements, i.e., webpage components that are likely to be cookie notices, and then we filter these elements.

Identifying candidate cookie notice elements. We employ two methods to identify candidate cookie notice elements, inspired by the works of Kampanos et al. [31] and Khandelwal et al. [32].

We first look for elements that match identifiers from the EasyList Cookie list, which is a crowd-sourced list used by adblockers to hide cookie notices. It contains a list of *selectors* that identify cookie notice HTML elements.

However, cookie notice detection based on this list results in both false positives and false negatives. False positives stem from websites where the selectors match multiple different elements that are not cookie notices. False negatives result from the list's incompleteness. For example, some websites have element identifiers that change constantly. For other websites, removing the cookie notice may break their functionality.

To detect more cookie notice candidates, we follow Khandelwal et al. [32] and inspect the *z-index* attribute of HTML elements, which indicates how elements are positioned with respect to each other. Since cookie notices are usually on top of other elements, we add elements with a positive z-index to our list of candidates.

Filtering detected elements. The list of candidates may include elements that do not correspond to cookie notices. To find a cookie notice element, we look for an element that contains the word "cookie" in a sentence that has a verb using a part-of-speech tagger. For this, we use language-specific spaCy models and language-specific variations of the word "cookie."

Note that Kampanos et al. [31] looked for elements that mention "cookie" without any additional filtering step. This may yield false positives when page footers contain "cookie policy" or "cookie preferences" links. We compare our method for cookie notice detection to both Kampanos et al. and Khandelwal et al. in Appendix A.

3.3 Cookie notice exploration

Once a cookie notice is detected, we explore its content by browsing its subpages, and we collect cookies after providing different types of consent, depending on the options given by the notice.

3.3.1 Text extraction

We view the cookie notice as a graph where each node corresponds to a state of the webpage that is reached by clicking on some sequence of interactive elements $e_1,\dots,e_i$. We extract the text of the cookie notice by exploring this graph using depth-first search (*DFS*).

After clicking on e_i, we have 3 possible scenarios:

- A new cookie notice element appears (e.g., after clicking on "Cookies Settings"). We use our cookie notice detection module to detect when this happens. In this case, we extract its text and its interactive elements and add them to the DFS stack of elements to be subsequently explored.

[2] The UK GDPR is almost identical to the EU GDPR and, in addition, some websites might still be obliged to apply the EU GDPR when visited from an EU IP address as we do.

[3] https://github.com/LibreTranslate/LibreTranslate

[4] https://github.com/Mimino666/langdetect

- Minor changes occur to the cookie notice element (e.g., we check a box or uncover more text). In this case, we add the newly discovered text to the set of extracted texts, if there is any.
- The cookie notice disappears, e.g., after clicking on "Accept" or "Close." In this case, we clear the browser's cache and cookies as this is crucial to ensure the cookie notice appears again when exploring further nodes. Websites usually store consent information in the browser to prevent the cookie notice from reappearing.

Since many elements only appear after clicking on a specific sequence of other elements, we define each node of the graph by the sequence of interactive elements $e_1, \ldots, e_i$ that lead to it. We explore such a node by reloading the website and clicking consecutively on the elements in the sequence.

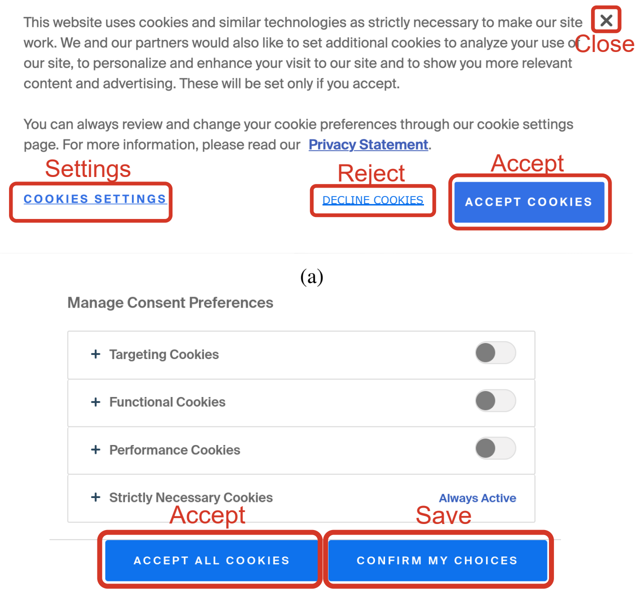
To extract interactive elements, we filter those with a non-negative `tabindex` attribute as these are accessible with tabbing. We also include DIV elements with a `role` attribute set to `button`, or a non-null `onclick` attribute. We use CSS selectors for element identification, which we retrieve using an open-source npm library.[5] For each explored node, we store newly uncovered text and translate non-English text to English using LibreTranslate.

Note that since the exploration tree can become very large, we set the maximum search depth to two and limit the maximum number of explored children per node to 50. For each node, if there are more than 50 interactive elements, we sort them in the order of their appearance in the HTML DOM and only keep the first 25 and last 25 elements. We assume that the skipped elements in the middle are repetitive in their content. A common case where we detect hundreds of interactive elements is when cookie notice subpages include information about all third-party vendors. Note that, in our experiments, the elements at the first layer of cookie notices are never truncated, whereas they are only truncated 3.2% of the time at the second layer.

3.3.2 Consent option detection and cookie extraction

We add structure to the cookie notice exploration using an interactive elements classification model, explained further in Section 4.2. Given the text of an interactive element, the model classifies it as: (1) a consent option among accept, reject, close, or save; (2) a settings button; or (3) other (negative sample). Fig. 2 shows an example of applying the interactive elements model to a cookie notice. We employ this model to detect consent options and associate them with the set of cookies stored by the websites when we interact with them.

If an element is classified as a consent option, we override this classification if it does not make the cookie notice disappear. This verification step increases the reliability of our method for consent options detection. It allows for filtering negative samples that are classified as consent options. It also allows for filtering elements with correctly classified text that correspond to options for individual settings, e.g., toggle switches, that users must interact with before clicking on a consent option that makes the cookie notice disappear.



This website uses cookies and similar technologies as strictly necessary to make our site work. We and our partners would also like to set additional cookies to analyze your use of our site, to personalize and enhance your visit to our site and to show you more relevant content and advertising. These will be set only if you accept.

You can always review and change your cookie preferences through our cookie settings page. For more information, please read our Privacy Statement.

(a)

(b)

Figure 2: Example of a cookie notice with annotated interactive elements. Fig. 2a illustrates the first layer of a cookie notice. Clicking on the "Cookies settings" button uncovers the second layer, shown in Fig. 2b. Both figures are annotated in red with labels of positive samples of the interactive elements classifier. Links, toggle switches, and other unannotated interactive elements correspond to negative samples.

For each filtered consent option, we extract the set of cookies resulting from interacting with it. After clicking on the element, we extract cookies following a strategy similar to Bollinger et al. [4]. We first browse the website by visiting random subpages—not including links in the cookie notice—and scrolling down to the bottom of each subpage. Then, we extract the cookies set with the *Cookie Instrument* provided by OpenWPM, which automatically collects cookies that are set by JavaScript or HTTP Responses. The browsing step is important. Urban et al. [51] showed that it increases the number of detected cookies by 36%.

In summary, we collect cookies in five independent crawling steps, depending on the available actions: accepting cookies, rejecting cookies, closing the notice without providing consent, saving the default choices, and not interacting with the notice. We infer from this information whether websites honor user choices as explained in Section 5.

[5] `https://github.com/autarc/optimal-select`

3.4 Deployment

To ensure that the websites crawled are under GDPR jurisdiction, we select websites that target EU or UK users and we crawl from an IP address located in the EU.

We considered websites from the Chrome User Experience Report [10] (*CrUX*). The dataset is based on anonymously collected user browsing histories. Although the dataset has potential biases as it only represents the data of users who opt-in to data collection, Ruth et al. [43] show that it represents website popularity significantly better than alternatives such as Alexa [1] or Tranco [37].

We collect websites with the popularity ranks 1k, 5k, and 10k, from the country-level CrUX datasets of 15 countries (see Table 3). This results in a crawling list of 97,090 websites. We crawl websites using 30 German datacenter IP addresses provided by The Bright Initiative from Bright Data [6]. We noticed no significant increase in the error rate when crawling using datacenter or ISP proxies.

We run the crawl on a consumer-grade PC with an AMD Ryzen 9 5950X CPU. This 16-core CPU allows us to run 30 OpenWPM browsers in parallel, achieving a crawling speed of 300 websites per hour. The crawl is not network-constrained and we only transfer 10 Mbit/s on average.

Ethical considerations We ignore `robots.txt` which we believe is meant for search engines and not research crawlers. As per the ethics committee at our institution, we do not require ethics approval for our crawling study since it does not involve human subjects. Moreover, our crawl does not harm website owners, since the computational costs are negligible and our findings are published in aggregated form without exposing individual websites. Regarding the legal aspect, we considered different legal regimes and concluded that our research does not violate laws such as fraud, trespass, or breach of contract since our intentions are to carry out good-faith privacy research.

4 Machine learning models

To automate the detection of violations and dark patterns, we develop machine learning models. These models are used to detect declared cookie purposes in the cookie notice text, to assist in crawl navigation using the interactive elements of the cookie notice, and to detect whether the website uses AA cookies appropriately. In this section, we present these models and how we trained them.

4.1 Declared purpose detection model

The GDPR requires websites to declare the purposes for data collection and processing thereby allowing the users to provide informed and explicit consent. Since most websites use plain language in their cookie notices to communicate these purposes, we developed a machine learning model that processes cookie notices and detects purpose declarations.

We use a dataset constructed by Santos et al. [45], who manually examined declared purposes and other legal requirements in a sample of cookie notices from the 1300 most visited English websites according to the Tranco ranking [37]. Their analysis is limited only to the first layer of cookie notices, but given the large variety of cookie notice designs, we expect the dataset to be rich enough to train models that generalize well to texts hidden in the subsequent cookie notice layers. Santos et al. [45] state that the annotation procedure of five researchers reached an agreement of 0.71-0.8 (Cohen's κ), which constitutes a substantial agreement according to Sim et al. [46]. However, given that the sample size is small, we simplify the task by merging the original labels into two groups as follows:

Purposes for Analytics/Advertising (*AA purpose*): profiling, advertising, custom content, analytics, and social media features.

Other purposes: essential functionalities, offering service, and website/UX enhancement.

We also pre-process the dataset by splitting its text into sentences using spaCy's pipeline and assigning labels to each sentence. We end up with 1171 sentences, 163 of which are labeled with an AA purpose.

We train a *purpose detection* BERT model [11] on this dataset. Given a sentence, this model predicts whether it mentions AA purposes. We train it with 5-fold cross-validation achieving an accuracy of 97.6% and an F1 score of 95.1%.

At test time, we split the text extracted during the crawl into sentences similar to our pre-processing step and apply the purpose detection model to each sentence. The website is considered to have a declared AA purpose if it is detected in at least one sentence.

4.2 Interactive elements model

We next build a model that allows the crawler to understand the functionality of the cookie notice elements, and, in particular, determine which elements likely represent consent options. We apply our text extraction method to crawl the top 12k Tranco websites with a `.uk` extension. This results in 2353 unique text samples extracted from interactive elements. We manually annotate each sample with one of the six labels: accept, reject, close, save, settings, and other. One of this paper's authors annotated the entire set, while another author validated 200 random samples, reaching an agreement of Cohen's $\kappa = 91\%$, which is an almost perfect agreement according to Sim et al. [46].

The accept, reject, close, and save labels correspond to consent options that the user may click on to make cookie notices disappear. The settings label corresponds to buttons that the user clicks on to uncover more consent options. The other label corresponds to negative samples that do not fall into any of these categories. An annotation example is shown in Fig. 2.

We train a BERT model on this dataset using 5-fold cross-validation. It achieves an accuracy of 95.1% and an F1 score of 90.9%.

4.3 Cookie classification model

We also build a model that decides for a set of cookies stored by the website, whether the website uses AA cookies. First, we use CookieBlock [4] to identify the purpose of each cookie in the set. CookieBlock is an ensemble that classifies cookies into one of four purposes, as defined by the UK ICC: necessary, functional, analytics, and advertising. The model relies on several features such as the cookie's entropy, its expiry date, and the edit distance between cookie updates. The features also include one-hot encoding of popular cookie names and domains. The model was trained on a dataset of cookies that was collected after crawling 27k websites implementing one of three CMPs—OneTrust, Cookiebot, and Termly—that provide labels for individual cookies. The dataset does not include CMP-specific cookies.

We re-train the model on a relabeled dataset where we merge the necessary/functional and analytics/advertising labels to match the categorization defined in Section 4.1. Then, we classify the website as *tracking* only if at least τ cookies were classified by CookieBlock as AA cookies. The threshold τ is needed for the unlikely case that a false positive occurs with the CookieBlock classification as we want to reduce the risk of falsely classifying websites as using AA cookies when they are not.

We evaluate the model on a dataset of 25k cookies collected by Bollinger et al. [4] from 3000 websites using the Cookiebot CMP, after opting out of consent. We train the model on the 245k cookies from the remaining 24k websites, collected after opting in to consent. The results are summarized in Table 1. We choose $\tau = 2$ where we achieve a precision of 98.7% and a recall of 91.8%. For our purposes, we would like our pipeline to produce a reliable lower-bound estimation of how many websites potentially violate privacy laws. We therefore aim to achieve high precision scores even at the expense of a lower recall, which justifies our choice of τ.

We additionally train the model on data from OneTrust and Termly only (14k websites, 137k cookies). We evaluate it on the same Cookiebot dataset and observe no significant difference in performance as shown in Table 1. This shows that the model is not biased towards the set of CMPs used for training, and it can therefore be used on any of the websites in our study.

	All CMPs			OneTrust+Termly		
τ	Precis.	Recall	Acc.	Precis.	Recall	Acc.
1	96.6	98.4	96.2	96.0	96.9	94.6
2	98.7	91.8	92.9	98.6	91.7	92.8
3	99.3	79.0	83.0	99.3	77.6	82.7

Table 1: Evaluation of cookie classification model. "All CMPs" include evaluation metrics of $g(\tau)$ trained on Cookiebot, OneTrust, or Termly websites. "OneTrust+Termly" include the same metrics for $g(\tau)$ trained on OneTrust and Termly websites. In both cases, the evaluation is performed on Cookiebot websites.

5 Observed violations and dark patterns

In this section, we review the legal requirements for cookie notices and describe the decision tree we implemented to automate the detection of potential privacy violations and dark patterns. We also present the results on our list of 97k websites sampled from the Chrome User Experience Report, as explained in Section 3.4. A summary decision tree for privacy violations is shown in Fig. 3, and a similar tree for dark patterns is shown in Fig. 5. Fig. 4 shows statistics on observed violations and dark patterns.

Legal background. The GDPR [19] states that any processing of personal information such as a name, an identification number, or location data, shall be lawful only if and to the extent that the data controller has a legal basis listed in Article 6 of GDPR. Except for narrowly defined cases, including Article 6(1)(b) of the GDPR which allows for the processing of personal information to fulfill a legal contract (interpreted narrowly in the Facebook case by the European Data Protection Board [16]), the only way to collect and process personal data for the purposes of tracking and personalized advertising is by a freely given, informed, explicit, and unambiguous consent of the data subject to the specific purposes (Article 7 and Recital 32). Under the ePrivacy Directive [18], most online marketing technologies and methods, including the use of cookies [2], require valid consent unless they are strictly necessary to provide the service (Article 5(3)).

The Data Protection Agencies (*DPA*) across EU member states apply the principles of the GDPR and ePrivacy Directive, although sometimes differing in their interpretation thereof. Recent attempts to harmonize rules on cookie notices across the EU have been more fruitful [17]. We focus our attention on potential violations that have been directly challenged by a DPA in a member state included in our sample or are clearly within the scope of the GDPR. Our findings can further be disaggregated to member states to account for potential differences between DPAs and national legal requirements, and we plan to explore these differences in future work.

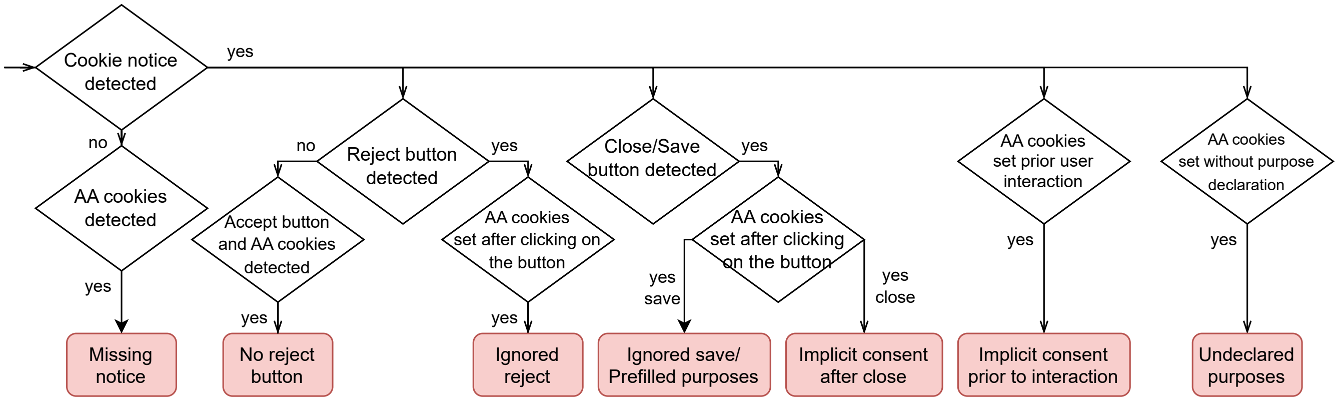

Cookie notice detected
yes
no
AA cookies detected
yes
Missing notice
Reject button detected
no
yes
Accept button and AA cookies detected
yes
No reject button
AA cookies set after clicking on the button
yes
Ignored reject
Close/Save button detected
yes
AA cookies set after clicking on the button
yes save
yes close
Ignored save/ Prefilled purposes
Implicit consent after close
AA cookies set prior user interaction
yes
Implicit consent prior to interaction
AA cookies set without purpose declaration
yes
Undeclared purposes

Figure 3: Decision tree for violations in cookie notices.

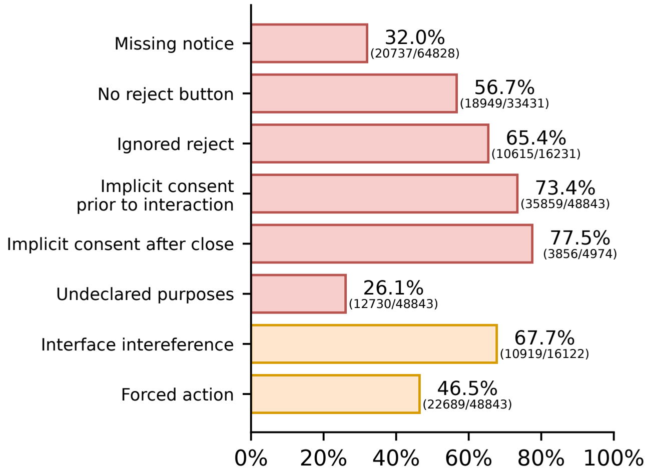


Figure 4: Number of websites with associated privacy violations and dark patterns.

We refer to court cases and guidelines in the text of individual violations and dark patterns, when applicable.

5.1 Observed violations

Out of a total of 97,090 websites, 92.8% were detected with a supported language. The language detection failed on 3.4% of the websites and 3.8% of the websites were detected with an unsupported language. We successfully crawled 85,443 websites, i.e., 94.8% of the websites with a supported language. We found that 72.2% of these websites contained at least one privacy violation. We summarize the violations as red bars in Fig. 4.

5.1.1 Missing cookie notice

A website's failure to meet the elementary obligation of obtaining consent from the data subject is most apparent in the case of websites that use AA cookies without a cookie notice. A missing notice constitutes a privacy violation across the EU. The French Data Protection Agency (*CNIL*) fined Amazon.fr 35M Euro, inter alia, for processing personal data without any prior notice [39].

Of the 64,828 websites where AA cookies were detected and most likely require consent, 32.0% were missing a cookie notice. Even if we account for a false positive rate of 21.9%, as shown in Section 7, 25.0% of websites are still missing cookie notices.

5.1.2 Effect of interactive elements

No reject button. For consent to be freely given, websites must offer a means to reject AA cookies. When the notice makes rejection more difficult or time-consuming than accepting cookies, it is considered coercion in some EU jurisdictions. In a recent decision, the CNIL fined Google and Facebook 150M and 90M Euro, respectively [40] for failing to provide a reject button when an accept button was present. These websites required users to navigate to settings in order to reject AA cookies. The same issue was deemed non-compliant by the European cookie banner taskforce [17]. We therefore report a missing reject button as a violation when the cookie notice contains an accept button without a matching reject button. We found this violation on 56.7% of the 33,431 websites with notices that included an accept button.

Ignored reject. When websites contain a reject button, our crawler inspects whether it correctly rejects all AA cookies. Using AA cookies after negative consent would require an alternative legal basis for data collection. Some websites, for example, claim that using these cookies is possible under the 'legitimate interest' legal basis of Article 6(1)(f) of the GDPR. However, the European cookie banner taskforce [17] has made it clear that websites can not circumvent the ePrivacy Directive consent requirement using GDPR legitimate interest. A court decision on this is pending [8].

We detect AA cookies after clicking on the reject button in 65.4% of the 16,231 websites that have a reject button in their cookie notice. Only 16.2% of these websites rely on legitimate interest. We believe that this violation results from a combination of ignorance and malice by the website owners and developers.

Implicit consent. The Court of Justice of the European Union (*CJEU*) ruling in the case of Planet49 [30] confirmed the German DPA interpretation of the GDPR and ePrivacy Directive that pre-checked checkboxes on consent banners are invalid forms of consent, apart from strictly necessary cookies. The recent taskforce report [17] confirmed this requirement and added that any form of inactivity (i.e., no interaction with the cookie notice) should not constitute consent under the GDPR or Article 5(3) of the ePrivacy Directive. We therefore study the use of AA cookies before users interact with the cookie notice. We denote this practice "Implicit consent prior to interaction" and detect it in 73.4% of the 48,843 websites that contain a cookie notice.

When a cookie notice has a close button, its functionality should not be the same as accepting cookies, because this violates the principle of positive and explicit consent as stated above. We call this an "Implicit consent after close" violation and detect it in 77.5% of the 4974 websites with a detected close button in their cookie notice.

5.1.3 Undeclared cookie purposes

One of the conditions for lawful consent under the GDPR (defined in Article 4(11)) is the specificity of the consent with respect to the declared purposes. When websites use cookies and other technologies for purposes that are not clearly mentioned in the notice provided to the user, such actions are likely to be deemed unlawful, as they are equivalent to acting without explicit consent.

We found that 26.1% of the 48,843 websites with cookie notices use AA cookies without declaring AA purposes in the initial text of the cookie notice. If we further explore the cookie notice, this number is reduced to 20.5%, since many websites list details on processing purposes in the cookie notice settings.

5.2 Observed dark patterns

The use of deceptive user interface designs to manipulate users' behaviors and choices is a murky area of law that has attracted attention in recent years. These practices are deployed to exploit, ultimately for profit, certain consumer traits. In the case of cookie notices, dark patterns have been found effective in nudging users [41]. The ability of online service providers to manipulate user behavior and choice, as well as extract consent for data sharing and tracking, has direct implications for consumer autonomy and welfare [21], and, therefore should be constantly monitored. Some of these practices could be considered illegal under the GDPR in specific jurisdictions [24, 54] and under the new Digital Services Act [20, Art. 25] for online platforms.

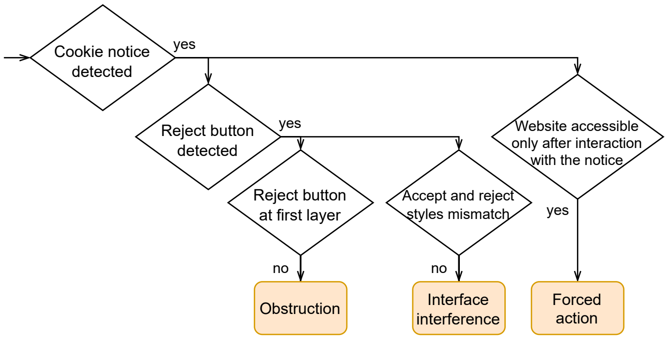

Cookie notice detected
yes
Reject button detected
yes
Website accessible only after interaction with the notice
Reject button at first layer
Accept and reject styles mismatch
yes
no
no
Obstruction
Interface interference
Forced action

Figure 5: Decision tree for dark patterns in cookie notices.

We implemented the detection of five dark patterns, specific to cookie notices, as defined by Soe et al. [48]. Below are our findings for two of these patterns. We skip the "Obstruction" pattern as it overlaps with the "No reject button" violation, which is a stricter observation. Similarly, "Sneaking" is covered by our analysis of legitimate interests in Section 5.1.2. Finally, we skipped "Nagging" as we rarely observed this pattern and it was not always reproducible. Our dark patterns observations are summarized by the orange bars in Fig. 4.

5.2.1 Interface interference

Previous studies [48, 49] defined interface interference as a 'design (color, font, size) of the buttons for accepting and rejecting cookies are not equivalent.' We analyze websites that have a *positive consent option* (an accept button) and a *negative consent option* (one of reject, save, or close buttons, where the first one is selected in that order) on the first layer of their cookie notice. For each button, we determine the dominant color with a k-means analysis of its screenshot and extract its text style (font, weight, and color). If we detect a mismatch in the colors; or if the text font or weight of the negative consent option is smaller than those of the positive consent option, then we report an "Interface interference" dark pattern. We consider two colors mismatched when their L_∞ distance is higher than $\frac{255}{4}$. We manually evaluate this threshold on 500 random websites and achieve 100% accuracy and we therefore do not employ more complex alternatives [33].

We detected "Interface interference" in 67.8% of the 16,122 websites that had both positive and negative consent options in the first layer of the cookie notice. This analysis assumes that nonequivalent interface elements are favoring the accept option, as it is impossible to pre-define what a user would find more attractive without studying the design language of each website in our sample.

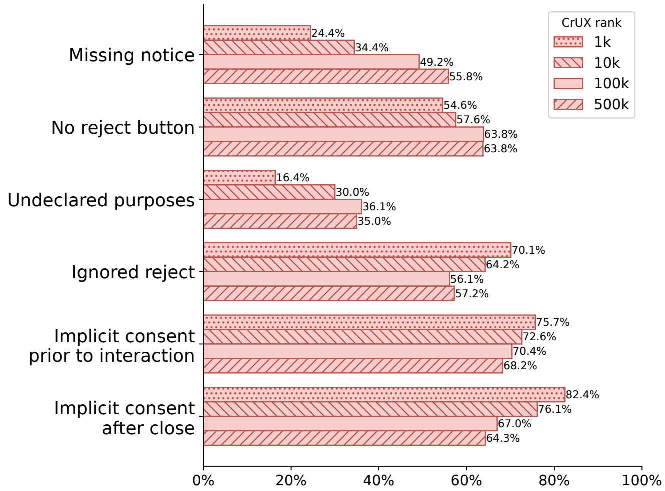


Figure 6: Ratio of websites with associated privacy violations, per rank.

5.2.2 Forced action

Forced action can appear in different forms, the harshest one being a "cookie wall" where a user can only enter the website after accepting all types of cookies and tracking technologies. Since such action might be considered illegal in some jurisdictions, websites might find appealing lighter forms of forced action to enhance acceptance rates. Previous studies [23, 48] have defined the forced action dark pattern as any requirement to engage with the cookie notice before entering the website.

For each website with a detected cookie notice, we sample five links not included in the cookie notice, and report a "Forced action" if we fail to click on each of them. This was detected in 46.4% of the 48,843 websites with a detected cookie notice.

We report "Forced action" as a dark pattern according to Gray et al. [23]. We however question whether this behavior is an intentional deception. Bakos et al. [3] showed that users fail to read legal notices, so requiring users to interact could raise their privacy awareness. To distinguish whether website designers act in good or bad faith, we tested this pattern in combination with other violations, but the results were inconclusive. Future studies could use a more fine-grained definition for this dark pattern and investigate the intentions more thoroughly.

5.3 Compliance by website popularity

The Chrome User Experience Report contains data on website popularity in each country allowing us to report more fine-grained results. We consider the popularity ranks (1k, 10k, 100k, and 500k)[6] and demonstrate that more popular websites

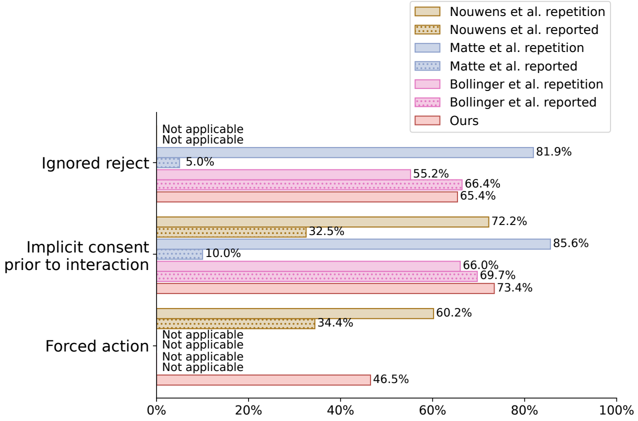


Figure 7: Comparison of our results with other studies.

are more likely to include a cookie notice and a reject button within the notice, as well as to list AA purposes. However, they also tend to ignore user reject actions or assume implicit consent. In other words, the popular websites keep a façade of compliance, while in reality, they harvest more user data than less popular websites [36].

We summarize observed violations depending on the popularity rank in Fig. 6. Notice that the majority of the pairwise observations are statistically significant. We use Fisher's exact test and apply the Holm–Bonferroni correction to the p-values, rejecting the hypothesis that results come from the same distribution when the p-value < 0.001. The significance of these results is also confirmed by the monotonic change in violations across ranks, with the exception of "Undeclared purposes" and "Ignored reject" in the 500k rank.

6 Bias analysis

In this section, we reproduce results from previous studies on potential violations that selected only websites employing certain CMPs [4, 38, 41]. We demonstrate how this selection criterion leads to a selection bias.

First, we compare the findings of each study to our findings on our broader collection of websites. Then, we reconstruct the selection of websites used in these studies using a combination of the Consent-O-Matic [28] CMP detection module and IAB Europe's Transparency & Consent Framework's API [15]. Fig. 7 summarizes our findings.

Comparison with Bollinger et al. Bollinger et al. [4] detect the "Implicit consent prior to interaction" violation in 69.7% of the 30k websites using one of the OneTrust, Cookiebot, or Termly CMPs. We detect the same violation in 66.0% of the 9434 websites with these CMPs compared to 73.4% on our broader collection of websites.

Bollinger et al. limit their analysis of the "Ignored reject"

[6] For this analysis, we crawled an additional 20k websites in the ranks 100k and 500k, sampled evenly across the countries selected in Section 3.4.

violation to websites using Cookiebot, which they detect in 66.4% of the 9.6k websites. We detect this violation in 55.2% of the 2323 Cookiebot websites with reject buttons and 65.4% on the broader collection of websites. This might indicate a slight improvement in compliance of Cookiebot from 66.4% to 55.2% over the past two years.

For each violation, Fisher's exact test rejects the hypothesis that the violation ratio on all websites and the violation ratio on the CMP-specific subset of websites come from the same distribution with p-values < 0.001 after applying the Holm–Bonferroni correction. This illustrates that the CMPs selected by Bollinger et al. show a higher level of compliance with consent requirements, despite not being fully compliant.

Comparison with Matte et al. Matte et al. [38] analyze the "Implicit consent prior to interaction" and "Ignored reject" violations on websites employing IAB Europe's TCF [14]. They detect these violations in 10% and 5% of the websites, respectively. Our analysis finds these violations in 85.6% of the 9132 IAB TCF websites and 81.9% of the 2548 IAB TCF websites with a reject button, respectively. This substantial discrepancy between these results could be explained by a difference in methodology. While we analyze all cookies set by the website to determine whether a website uses AA cookies, Matte et al. investigate whether IAB TCF websites use a particular cookie (e.g., `consent_cookie`) to store the user's consent choice (e.g., `tracking_allowed=True`). Their measurements therefore do not correspond to actual cookie usage.

Furthermore, we note a significant increase in detection ratios for these violations compared to our results on the broader collection of websites in Fig. 7. These results are statistically significant, with p-values < 0.001 using the same test as described above. Despite the strict evaluation criteria by Matte et al., their sample induced a bias in the observations.

IAB was also the focus of the Belgian DPA, as they questioned IAB's compliance with the GDPR, specifically the permitted use of legitimate interest to track users [12, 53]. Our findings could be capturing the fact that several CMPs stopped supporting IAB TCF as a result of the legal uncertainty, leaving risk-seeking CMPs in our sample.

Comparison with Nouwens et al. Nouwens et al. [41] analyze the "Implicit consent prior to interaction" violation and the "Forced action" dark pattern, which they find respectively in 32.5% and 34.4% of 680 UK websites employing a CMP among Cookiebot, Crownpeak, OneTrust, Quantcast, and Trustarc.

Our analysis found these behaviors in 72.2% and 60.2% of the 12.3k websites with these CMPs. The discrepancy with the results of Nouwens et al. could be attributed to the small number of websites covered by their study, which likely does not accurately represent all websites using the selected CMPs. The discrepancy in the implicit consent reporting could also be attributed to a difference in methodology. They rely only on cookie notices, while we also consider actual cookie usage.

The "Implicit consent prior to interaction" violation is observed with a similar ratio (73.4%) on all websites. A selection of multiple CMPs has averaged out the results for this particular violation. We observe however a disparity for the "Forced action" dark pattern that we find in 46.5% of websites from the broader crawling list. We also observe significant disparities for violations such as "No reject button" and "Undeclared purposes," again with p-values < 0.001.

Violation/Dark pattern	Precision	Recall
Missing notice	78.1	94.3
No reject button	91.5	65.2
Ignored reject	90.7	62.8
Implicit consent after close	94.7	58.1
Implicit consent prior to interaction	91.2	86.5
Undeclared purposes	97.5	90.6
Interface interference	100.0	72.0
Forced action	100.0	83.8

Table 2: Evaluation of the violation and dark pattern detection on 500 websites sampled randomly from the crawling list.

7 Manual evaluation

Our methods for violation and dark pattern detection depend on machine learning models and heuristics, which are prone to false positives and negatives. To assess the performance of our methods, we manually annotated 500 websites randomly sampled from the crawling list. To manually determine whether a website uses AA cookies, we rely on the platforms Cookiepedia [42] and Cookiedatabase.org [7], which include purpose annotations of cookies provided by human experts. Note that these databases are skewed to classify cookies as necessary [4] while our AA cookies detection might be correct. We skip cookies that do not have matching annotations in these platforms to provide websites further benefit of the doubt. For these reasons, the evaluation below is strict and should serve as a lower bound on the performance of our methods. The results of the manual evaluation are summarized in Table 2.

Missing cookie notice We identify a "Missing cookie notice" violation when AA cookies are detected while a cookie notice is not detected. The violation was correctly detected in 82 out of 87 websites, and falsely detected in 24 websites. This results in a precision of 78.1% and a recall of 94.3%. The false positives mainly stem from false negatives in the cookie notice detection (see Appendix A). We expect the precision to improve as the spaCy models and EasyList improve over time, and particularly as the diversity of cookie notices diminishes with the CMP market unification and increased web development outsourcing.

No reject button The "No reject button" violation was correctly detected in 107 out of 164 websites (recall 65.2%) and falsely detected in 10 websites (precision 91.5%). Note that the low recall stems from our conservative approach to reduce false positives by only identifying a "No reject button" violation if an accept button is detected. Without this restriction, we would have a recall of 88.2% and a precision of 85.6%.

Ignored reject The "Ignored reject" violation was correctly detected in 49 out of 78 websites (recall 62.8%). It was incorrectly identified in 5 websites (precision 90.7%). These false positives include one website where a close button was incorrectly detected as a reject button. The remaining four websites had their reject button correctly identified but did not use AA cookies according to Cookiepedia and Cookiedatabase.org.

Implicit consent The "Implicit consent after close" violation was correctly detected in 18 out of 31 websites (recall 58.1%) with one false positive (precision 94.7%). The "Implicit consent prior to interaction" violation was correctly detected in 186 out of 215 websites (recall 86.5%) and incorrectly detected in 18 websites (precision 91.2%). The false negatives are mainly due to undetected cookie notices or undetected buttons, while the false positives are due to AA cookies misclassification.

Undeclared purposes The "undeclared purpose" violation was correctly detected in 77 out of 85 websites (recall 90.6%), with two false positives (precision 97.5%). Six false negatives resulted from undetected cookie notices. The other two false negatives and the two false positives correspond to misclassifications by the purpose detection model. Overall, the purpose detection model achieves a good performance despite the small number of positive samples in the training set.

Dark patterns "Interface interference" and "Forced action" were correctly detected in 60 out of 83 websites (recall 72%), and 134 out of 160 websites (recall 83.8%), respectively. The false negatives were mainly due to undetected cookie notices or undetected buttons. Both dark patterns were detected with 100% precision.

8 Limitations

8.1 Translation inaccuracy

Since our work relies on the processing of natural language rather than cookie notice APIs, we require machine translation to support multiple languages. We observed however that LibreTranslate has trouble translating the short texts of interactive elements in some languages. For example, the Greek version of "Accept all cookies" translates to "Cookie policy" in English. In this work, our solution depended on the frequency of such mistakes. In languages where mistranslations were rare, like German and Danish, we manually fixed these wrongly translated texts. However, we removed Greek websites completely, since the model performs poorly on the majority of inputs. In the future, models will only improve and such manual edits may no longer be necessary. Alternatively, one can use better translation services such as Google Translate or DeepL Translator, but their costs scale poorly with crawling tens of thousands of websites. Nevertheless, non-English speaking websites are understudied. We therefore consider their inclusion, even using imperfect translation, as a significant added value to the generalizability of our study.

8.2 False violation reporting

We emphasize that the violations observed by our automated procedure cannot be directly taken by a court or DPA to enforce fines. Given the risks of false positives, one must inspect and confirm them manually. We describe possible reasons for false reporting by our procedure.

False positives. Our privacy violation and dark pattern detection is dependent on machine learning models and heuristics. The input of these models and heuristics are natural language, cookie content, and cookie notice interface visuals, all of which are prone to ambiguous interpretation. Our procedure can therefore falsely report violations.

We addressed this risk as follows. First, we train the models on a limited number of classes to observe the most crucial aspect of whether data processing purpose requires consent or not. Multi-label models would give more details on violation types at the cost of more false positives. Second, since the cookie prediction model of Bollinger et al. [4] does not achieve sufficient accuracy, we require observing multiple AA cookies. This gives websites the benefit of the doubt and restricts our observations only to websites that violate requirements more seriously by having a larger number of AA cookies.

We also evaluated the false positive rates on a set of 500 websites. We observed average false positive rates of 9.4% and 0.0% for privacy violations and dark patterns, respectively. Overall, we consider these rates to be low enough for our observations to be representative of the compliance picture.

Ambiguity. Many cookie notices use vague statements such as "This website uses cookies to improve user experience." The dataset by Santos et al. classifies these purposes as Website/UX enhancement. However, as mentioned above, we decided to simplify the problem to a binary classification.

Our decision to classify websites with this ambiguous statement as non-AA purpose results in reporting them as violators if they use AA cookies. We argue that even if this text declares the cookies in use, the consent is still invalid since it is ambiguous [19, Art. 7 and Rec. 32].

Regulation applicability. We base the consent requirements on EU privacy laws, mostly the GDPR. While the websites in our dataset are located all over the world, our reliance on the CrUX data ensures that a significant portion of users come from countries where the GDPR is applicable. If these websites target a significant portion of EU citizens and profit from this user base, they are also required to follow the GDPR [19, Art. 3]. When the CrUX data is imprecise, the number of website visitors from the EU could be checked using other paid sources, such as Similarweb [47].

8.3 Bias

One of our goals is to address the bias present in prior studies. However, our methods are themselves still prone to some biases. First, the CrUX list is based on only Chrome users who participate in the data collection. It is likely that these users are less privacy-aware and might visit different websites than average users. We used the CrUX list as recommended by Ruth et al. [43], who showed that biases related to crawling lists can only be measured by large Internet companies. Second, our traffic originates in Germany. While it is possible that websites treat German users differently from those from other EU countries, we expect such behavior to be rare, and we have not evaluated this due to computational requirements. Third, Degeling et al. [9] and Kampanos et al. [31] find that different countries can show different trends. Therefore, our country selection could cause a bias that can be addressed given translation resources for the missing languages.

8.4 Adversarial modifications

We did not address websites that might modify their cookie notice and cookie content to counterattack our models. Prior to the publication of this work, our NLP models were closed-source, so websites could not modify their notices to evade our detection. Our reported results should therefore be free of adversarial modifications. Whether websites will later modify their cookie notices to fool our detection classification methods is a topic for future work. Observing websites designing notices deceptive not only to users (dark patterns) but also to automated methods like ours would be an interesting legal case study on whether it is an intentional infringement according to Art. 83 of the GDPR.

9 Related work

Several studies have focused on websites' non-compliance with privacy regulations. We restrict our comparison to studies of cookie notices. We also refer readers to a meta-study by Kretschmer et al. [34], which summarizes empirical publications measuring GDPR impact on cookie notices and privacy policies.

Long-term studies Hils et al. [26] studied the emergence of CMPs and their influence on data collection consent from January 2018 to September 2020. They measured the popularity of six major CMPs and their market competition in the context of court decisions, fines, and new laws. They also surveyed the defined data processing purposes assigned to vendors and observed that accepting cookies takes less time than declining them. Degeling et al. [9] studied the adoption of cookie notices after the GDPR was enacted, finding that the presence of cookie notices increased by a third from January to May 2018. They also manually inspected what choices the notices offer, finding that the vast majority provided either no option or only an accept-all option.

Trevisan et al. [50] measured implicit consent prior to interaction with cookie notice. They detect websites' usage of AA cookies when at least one 3rd-party cookie is created by a domain in Ghostery or Disconnect advertising lists. Their AA cookie detection is therefore prone to falsely flagging advertiser's cookies that remain empty until the user consents to tracking, as many vendors do. Our work does not have this limitation and it can also detect 1st-party trackers. Trevisan et al. sampled websites from Similarweb, taking the top 100 websites of 25 EU countries and 25 categories, totaling almost 36k websites. They showed that 49% of the websites use such cookies prior to consent. They also investigated violations on a smaller sample over the period of four years, finding negligible differences in compliance. In a follow-up work by similar authors, Jho et al. [29] studied websites after accepting all cookies. They implemented an accept-button detector based on a keyword search with 95% accuracy. Our interactive element model achieves a similar accuracy for multi-class classification, allowing us to inspect cookie usage not only after accepting all cookies but also after taking other actions.

CMPs and Transparency & Consent Framework (*TCF*) Bollinger et al. [4] studied the compliance of websites with the OneTrust, Cookiebot, and Termly CMPs. They detected the presence of these CMPs on almost 27k websites from the 7M Tranco list. They reported eight types of violations that were observable given the data that these specific CMPs contained, observing that only 5.3% of websites were compliant with these requirements. Note that most violations reported in their work require websites to declare detailed per-cookie information, which is only provided by a few CMPs and is not required by privacy regulations. We do not include these violations as we would like our methods to be applicable on websites independent of CMP implementations.

Nouwens et al. [41] investigated consent on 680 websites of the top 10k UK Alexa list that use one of the Cookiebot, Cronwpeak, OneTrust, QuantCast, or TrustArc CMPs. They observe implicit consent, forced action, and pre-checked options, finding that only 11.8% of the websites are compliant.

Matte et al. [38] focused on websites with a CMP implementing IAB TCF. They observed the "Implicit consent prior to interaction" violation in 10% of the websites and the "Ignored reject" violation in 5% of the websites. We compare our results to all three studies and analyze their selection bias in Section 6.

Opt-out functionality Sanchez-Rola et al. [44] manually inspected cookie notice options shown to users from the EU, US, and China. Surprisingly, they found that cookie notices shown to EU users had fewer options and that the "ignored reject" violation was observed more in the EU than in the US or China. This analysis is complementary to our work.

Khandelwal et al. [32] developed an extension that automatically fills out cookie notices. They relied on two machine learning models. The first model determines, given an HTML element's text, whether it corresponds to a cookie notice. The second model takes as input the cookie notice in a machine-readable format and predicts the set of actions needed to disable non-essential cookies. They also analyze cookie notices on the top 100k Tranco websites, showing that they detect cookie notices in 52.7% of the websites, that 35.4% of the notices had multiple layers, and that 21.5% of the notices include a one-click opt-out option. In their analysis, Khandelwal et al. neither studied compliance with respect to privacy regulations nor did they condition their analysis on cookie usage or declared purposes as we do.

Dark patterns Soe et al. [48] studied five dark patterns on 300 news websites popular with EU users. They found that only three websites do not use any of the dark patterns they consider. Our dark patterns methods used the definitions from their work, but we observed "Interface interference" and "Forced action" patterns more often than Soe et al. This might be caused by the low reliability of the manual work that they report (inter-annotator agreement of 0.67).

Krisam et al. [35] studied popular German websites, finding that 85% of the websites nudge users to accept all cookies and that only 21.5% of the websites contain a reject button. We observed significantly more reject buttons, likely as a result of the fines imposed on Google and Facebook for this violation [40].

Kampanos et al. [31] used a crawler that detects cookie notices using EasyList and detects consent choices. They crawled 17k top websites from the UK and Greece and found cookie notices in 44% of websites and an opt-out option in only 6% of the websites. However, their analysis only focuses on the first layer of cookie notices. Moreover, the detection of consent options is keyword-based, which is only able to detect 56.7% of accept buttons and 33.4% of reject buttons found by our models.

10 Conclusion

We have developed and applied machine learning methods to the first general, automated, large-scale analysis of cookie notice compliance. The resulting system interacts with a wide variety of intricate cookie notices, identifying their interactive elements, determining which cookies are declared, and retrieving the cookies used by the website.

Thanks to our method, we have analyzed a much larger and more representative sample of the websites within the GDPR jurisdiction than previous studies, overcoming their limitations. As a result, we have provided corrected statistics for previous studies, which suffered from a selection bias, and new statistics for certain types of violations and dark patterns. Even after these corrections, we still observe a significant prevalence of potential violations. Indeed, 65.4% of websites ignore user choices when they explicitly reject consent and 72.2% of websites violate at least one of the requirements checked by our methods. Furthermore, we discovered that more popular websites are more likely to ignore user choices despite having more compliant cookie notice interfaces. This highlights the need for more enforcement to catch privacy violations, which automated methods such as ours can foster.

As future work, we plan to investigate the influence of individual CMPs and the crawl's IP location on website compliance, utilizing our rich multilingual sample. For example, we aim to determine whether websites paying for CMPs comply more or less than those websites using free services. We also plan to monitor compliance over time, measuring the effects of new regulations (like the Digital Marketing Act, the Digital Services Act, and the planned ePrivacy Regulation), new court decisions, and new privacy-preserving technologies.

Acknowledgment

We thank the Bright Initiative [6] for giving us access to their proxy services and the reviewers for their recommendations.

References

[1] Alexa top million. https://www.alexa.com/; Last accessed on: 2023.10.01.

[2] Article 29 Working Party, European Union. Guidelines on consent under regulation 2016/679 (WP259 rev.01), 2018. https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=623051; Last accessed on: 2023.09.30.

[3] Yannis Bakos, Florencia Marotta-Wurgler, and David R Trossen. Does anyone read the fine print? Consumer attention to standard-form contracts. *The Journal of Legal Studies*, 43(1):1–35, 2014.

[4] Dino Bollinger, Karel Kubicek, Carlos Cotrini, and David Basin. Automating cookie consent and GDPR violation detection. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 2893–2910, Boston, MA, August 2022. USENIX Association.

[5] C. Bösch, B. Erb, F. Kargl, Henning Kopp, and Stefan Pfattheicher. Tales from the dark side: Privacy dark strategies and privacy dark patterns. *Proceedings on Privacy Enhancing Technologies*, 2016:237 – 254, 2016.

[6] Bright Data. The bright initiative. https://brightinitiative.com/; Last accessed on 2023.08.08, 8 2023.

[7] Cookiedatabase.org. https://cookiedatabase.org/ Last accessed on: 2023.07.11.

[8] Court of Justice of the European Union. Request for a preliminary ruling case C-17/22, 2022. https://curia.europa.eu/juris/showPdf.jsf?text=&docid=255645&pageIndex=0&doclang=EN&mode=lst&dir=&occ=first&part=1&cid=6693245; Last accessed on: 2023.02.02.

[9] Martin Degeling, Christine Utz, Christopher Lentzsch, Henry Hosseini, Florian Schaub, and Thorsten Holz. We value your privacy... now take some cookies: Measuring the GDPR's impact on web privacy. *CoRR*, abs/1808.05096, 2018.

[10] Google Developers. Chrome User Experience Report. https://web.archive.org/web/20230123160339/https://developer.chrome.com/docs/crux/, January 2023.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[12] Belgian DPA. Decision on the merits 21/2022 of 2 February 2022, unofficial translation from Dutch, case number DOS-2019-01377, February 2022. https://edpb.europa.eu/system/files/2022-03/be_2022-02_decisionpublic_0.pdf; Accessed on: 2023.02.06.

[13] Steven Englehardt and Arvind Narayanan. Online tracking: A 1-million-site measurement and analysis. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, page 1388–1401, New York, NY, USA, 2016. Association for Computing Machinery.

[14] IAB Europe. IAB Europe Transparency and consent framework policies, May 2021. https://iabeurope.eu/iab-europe-transparency-consent-framework-policies/; Last accessed on 2022.12.30.

[15] IAB Europe. IAB Europe Transparency and consent framework tool suite, 2022. https://github.com/InteractiveAdvertisingBureau/iabtcf-es; Last accessed on: 2022.12.30.

[16] European Data Protection Board. Facebook and Instagram decisions: "Important impact on use of personal data for behavioural advertising", 2023. https://edpb.europa.eu/news/news/2023/facebook-and-instagram-decisions-important-impact-use-personal-data-behavioural_en; Last accessed on: 2023.02.03.

[17] European Data Protection Board. Report of the work undertaken by the Cookie Banner Taskforce, 2023. https://edpb.europa.eu/our-work-tools/our-documents/report/report-work-undertaken-cookie-banner-taskforce_en; Last accessed on: 2023.01.30.

[18] European Parliament, Council of the European Union. Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications), July 2002. http://data.europa.eu/eli/dir/2002/58/oj; Last accessed on: 2021.02.06.

[19] European Parliament, Council of the European Union. Regulation (EU) 2016/679 Of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), April 2016. http://data.europa.eu/eli/reg/2016/679/2016-05-04; Last accessed on: 2023.02.06.

[20] European Parliament, Council of the European Union. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act), October 2022. https://eur-lex.europa.eu/eli/reg/2022/2065/oj; Last accessed on: 2023.02.06.

[21] Ariel Ezrachi and Maurice E Stucke. Digital platforms inhibit innovation to address today's most pressing issues. *USApp–American Politics and Policy Blog*, 2022.

[22] Paul Grassl, Hanna Schraffenberger, Frederik Zuiderveen Borgesius, and Moniek Buijzen. Dark and bright patterns in cookie consent requests. *PsyArXiv*, 2020.

[23] Colin M Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L Toombs. The dark (patterns) side of UX design. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, CHI '18, pages 1–14, New York, NY, USA, 2018. Association for Computing Machinery.

[24] Colin M Gray, Cristiana Santos, Nataliia Bielova, Michael Toth, and Damian Clifford. Dark patterns and the legal requirements of consent banners: An interaction criticism perspective. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2021.

[25] Philip Hausner and Michael Gertz. Dark patterns in the interaction with cookie banners. *arXiv preprint arXiv:2103.14956*, 2021.

[26] Maximilian Hils, Daniel W. Woods, and Rainer Böhme. Measuring the emergence of consent management on the web. In *Proceedings of the ACM Internet Measurement Conference*, IMC '20, page 317–332, New York, NY, USA, 2020. Association for Computing Machinery.

[27] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. 2017.

[28] Rolf Bagge Janus Bager Kristensen. Consent-O-Matic, 2020. https://github.com/cavi-au/Consent-O-Matic; Last accessed on 2023.09.25.

[29] Nikhil Jha, Martino Trevisan, Luca Vassio, and Marco Mellia. The internet with privacy policies: Measuring the web upon consent. *ACM Transactions on the Web (TWEB)*, 16(3):1–24, 2022.

[30] Judgement of the Court (Grand Chamber). Case C-673/17 Planet49 GmbH v Bundesverband der Verbraucherzentralen und Verbraucherverbände – Verbraucherzentrale Bundesverband e.V. ECLI:EU:C:2019:246, 1 October 2019. http://curia.europa.eu/juris/document/document.jsf?docid=218462&doclang=EN; Last accessed on: 2022.12.30.

[31] Georgios Kampanos and Siamak F Shahandashti. Accept all: The landscape of cookie banners in Greece and the UK. In *IFIP International Conference on ICT Systems Security and Privacy Protection*, page to appear. Springer, 2021.

[32] Rishabh Khandelwal, Asmit Nayak, Hamza Harkous, and Kassem Fawaz. Automated cookie notice analysis and enforcement. In *32st USENIX Security Symposium (USENIX Security 23)*, Anaheim, CA, August 2023. USENIX Association.

[33] Simon Koch, Benjamin Altpeter, and Martin Johns. The OK is not enough: A large scale study of consent dialogs in smartphone applications. In *32st USENIX Security Symposium (USENIX Security 23)*, Anaheim, CA, August 2023. USENIX Association.

[34] Michael Kretschmer, Jan Pennekamp, and Klaus Wehrle. Cookie banners and privacy policies: Measuring the impact of the GDPR on the web. *ACM Transactions on the Web (TWEB)*, 15(4):1–42, 2021.

[35] Chiara Krisam, Heike Dietmann, Melanie Volkamer, and Oksana Kulyk. Dark patterns in the wild: Review of cookie disclaimer designs on top 500 German websites. In *Proceedings of the 2021 European Symposium on Usable Security*, pages 1–8, 2021.

[36] Filippo Lancieri. Narrowing data protection's enforcement gap. *Maine Law Review*, 74:15, 2022.

[37] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, and Wouter Joosen. "Tranco: A research-oriented top sites ranking hardened against manipulation". In *Proceedings of the 26th Annual Network and Distributed System Security Symposium*, NDSS 2019, February 2019.

[38] C. Matte, N. Bielova, and C. Santos. Do cookie banners respect my choice? Measuring legal compliance of banners from IAB Europe's Transparency and consent framework. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 791–809, 2020.

[39] Commission nationale de l'informatique et des libertés (CNIL). Délibération SAN-2020-013 du 7 décembre 2020, 2020. https://www.legifrance.gouv.fr/cnil/id/CNILTEXT000042635729; Last accessed on: 2023.02.01.

[40] Commission nationale de l'informatique et des libertés (CNIL). Cookies: the CNIL fines Google a total of 150 million euros and Facebook 60 million euros for non-compliance with French legislation, January 2022. https://www.cnil.fr/en/cookies-cnil-fines-google-total-150-million-euros-and-facebook-60-million-euros-non-compliance; Last accessed on: 2023.01.30.

[41] Midas Nouwens, Ilaria Liccardi, Michael Veale, David Karger, and Lalana Kagal. Dark patterns after the gdpr:

Scraping consent pop-ups and demonstrating their influence. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–13, 2020.

[42] OneTrust. Cookiepedia. https://cookiepedia.co.uk/; Last accessed on 2023.07.11.

[43] Kimberly Ruth, Deepak Kumar, Brandon Wang, Luke Valenta, and Zakir Durumeric. Toppling top lists: Evaluating the accuracy of popular website lists. In *Proceedings of the 22nd ACM Internet Measurement Conference*, IMC '22, page 374–387, New York, NY, USA, 2022. Association for Computing Machinery.

[44] Iskander Sanchez-Rola, Matteo Dell'Amico, Platon Kotzias, Davide Balzarotti, Leyla Bilge, Pierre-Antoine Vervier, and Igor Santos. "Can I opt out yet?": GDPR and the global illusion of cookie control. In *Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security*, Asia CCS '19, page 340–351, New York, NY, USA, 2019. Association for Computing Machinery.

[45] Cristiana Santos, Arianna Rossi, Lorena Sánchez Chamorro, Kerstin Bongard-Blanchy, and Ruba Abu-Salma. Cookie banners, what's the purpose? Analyzing cookie banner text through a legal lens. *CoRR*, abs/2110.02597, 2021.

[46] Julius Sim and Chris C Wright. The Kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy*, 85(3):257–268, 03 2005.

[47] Similarweb. Website traffic – check and analyze any website. https://www.similarweb.com/; Last accessed on: 2023.10.01.

[48] Than Htut Soe, Oda Elise Nordberg, Frode Guribye, and Marija Slavkovik. Circumvention by design-dark patterns in cookie consent for online news outlets. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*, pages 1–12, 2020.

[49] Alina Stöver, Nina Gerber, Christin Cornel, Mona Henz, Karola Marky, Verena Zimmermann, and Joachim Vogt. Website operators are not the enemy either - Analyzing options for creating cookie consent notices without dark patterns. In Karola Marky, Uwe Grünefeld, and Thomas Kosch, editors, *Mensch und Computer 2022 - Workshopband*, Bonn, 2022. Gesellschaft für Informatik e.V.

[50] Martino Trevisan, Stefano Traverso, Eleonora Bassi, and Marco Mellia. 4 years of EU cookie law: Results and lessons learned. *Proceedings on Privacy Enhancing Technologies*, 2019:126–145, 04 2019.

[51] Tobias Urban, Martin Degeling, Thorsten Holz, and Norbert Pohlmann. Beyond the front page: Measuring third party dynamics in the field. In *Proceedings of The Web Conference 2020*, page 1275–1286, New York, NY, USA, 2020. Association for Computing Machinery.

[52] Christine Utz, Martin Degeling, Sascha Fahl, Florian Schaub, and Thorsten Holz. (Un)informed consent: Studying GDPR consent notices in the field. *CoRR*, abs/1909.02638, 2019.

[53] Michael Veale, Midas Nouwens, and Cristiana Santos. Impossible asks: Can the Transparency and Consent Framework ever authorise real-time bidding after the Belgian DPA decision? *Technology and Regulation*, pages 12–22, 2022.

[54] Amit Zac, Yun-Chun Huang, Amédée von Moltke, Christopher Decker, and Ariel Ezrachi. Dark patterns and online consumer vulnerability. *Available at SSRN 4547964*, 2023.

A Cookie notice detection evaluation

We evaluate the cookie notice detection on the same 500 websites on which we performed the manual evaluation (Section 7). We achieve 100.0% precision and 86.9% recall.

As baselines, we compare to the cookie notice detection implemented by Khandelwal et al. [32] and Kampanos et al. [31]. Khandelwal et al. detect cookie notices using a BERT model that classifies the text of HTML elements with a positive z-index. Additionally, they consider the first 3 and last three HTML elements. We collect a dataset on which we train a similar BERT model following the steps described in their paper. We translate non-English text to English using LibreTranslate before feeding it to the BERT model. We achieve 97.4% precision and 78.2% recall with this method.

Kampanos et al. [31] rely on EasyList selectors to select candidate cookie notice elements. They filter them solely based on the presence of the word "cookie." We achieve 98.8% precision and 48.1% recall with this method.

Our cookie notice detection method therefore outperforms methods from prior studies in both precision and recall. Our use of both the z-index and the EasyList selectors to identify candidate cookie notice elements allows us to detect more cookie notices than these methods. We also avoid false positives, which correspond to page footers containing the word "cookie". Moreover, our method is significantly faster than Khandelwal et al.'s method since they depend on a BERT model that is more computationally expensive than the spaCy models we use.

Language	Countries	LT	spaCy	Us
Bulgarian	BG			
Croatian	HR			
Czech	CZ	✓		
Danish	DK	✓	✓	✓
Dutch	BE, LU, NL	✓	✓	✓
English	IE	✓	✓	✓
Estonian	EE			
Finnish	FI	✓	✓	✓
French	BE, FR, LU	✓	✓	✓
German	AT, BE, DE	✓	✓	✓
Greek	GR	✓	✓	
Hungarian	HU	✓		
Irish	IE	✓		
Italian	IT	✓	✓	✓
Latvian	LV			
Lithuanian	LT		✓	
Maltese	MT			
Portuguese	PT	✓	✓	✓
Polish	PL	✓	✓	✓
Romanian	RO		✓	
Slovak	SK	✓		
Slovenian	SI			
Spanish	ES	✓	✓	✓
Swedish	SE	✓	✓	✓

Table 3: List of common languages in the EU along with the countries where they are official. For each language we indicate whether it is supported by LibreTranslate (LT), spaCy, and our crawler (Us).

B Violations

Ignored save/Prefilled purposes. The Court of Justice of the European Union (CJEU) ruling in the case of [30] confirmed the German DPA interpretation of the GDPR and ePrivacy Directive that pre-checked checkboxes on consent banners are invalid forms of consent, apart from strictly necessary cookies. The taskforce report confirmed, on the European level, that pre-ticked boxes to opt-in do not lead to valid consent, as well as that inactivity (i.e., without positive action by the user) should not constitute consent under the GDPR or Article 5(3) of the ePrivacy Directive [17].

We report an "Ignored save/Prefilled purposes" violation when AA cookies are detected after clicking on "Save." We observe this behavior in 69.0% of the 12,951 websites with a detected save button. We achieve a precision of 91.1% and a recall of 66.1% on the manually annotated dataset introduced in Section 7.