ETH zürich

Karel Kubíček's site

Automating Cookie Consent and GDPR Violation Detection

Authors: Dino Bollinger, Karel Kubicek karel.kubicek@inf.ethz.ch, Carlos Cotrini, David Basin

Abstract: *The European Union's General Data Protection Regulation (GDPR) requires websites to inform users about personal data collection and request consent for cookies. Yet the majority of websites do not give users any choices, and others attempt to deceive them into accepting all cookies. We document the severity of this situation through an analysis of potential GDPR violations in cookie banners in almost 30k websites. We identify six novel violation types, such as incorrect category assignments and misleading expiration times, and we find at least one potential violation in a surprising 94.7% of the analyzed websites.*

We address this issue by giving users the power to protect their privacy. We develop a browser extension, called CookieBlock, that uses machine learning to enforce GDPR cookie consent at the client. It automatically categorizes cookies by usage purpose using only the information provided in the cookie itself. At a mean validation accuracy of 84.4%, our model attains a prediction quality competitive with expert knowledge in the field. Additionally, our approach differs from prior work by not relying on the cooperation of websites themselves. We empirically evaluate CookieBlock on a set of 100 randomly sampled websites, on which it filters roughly 90% of the privacy-invasive cookies without significantly impairing website functionality.

- Conference page: Usenix Security 2022
- Download author pre-print of the paper: PDF
- Download extended version of paper (Dino Bollinger's Master's thesis): PDF
- Presentation: Google slides
- See full conference talk: YouTube

- SOUPS poster **CookieBlock & CookieAudit: Fixing Cookie Consent with ML**: pdf, BibTeX, and page
- Invited talk: slides
- Download artifact that won distinguish artifact award: datasets, crawler, classifier, and violation detection scripts
- Try our extension CookieBlock: Chrome, Firefox, Edge, Opera, or check out the source code
- Give us feedback on the extension

BibTeX

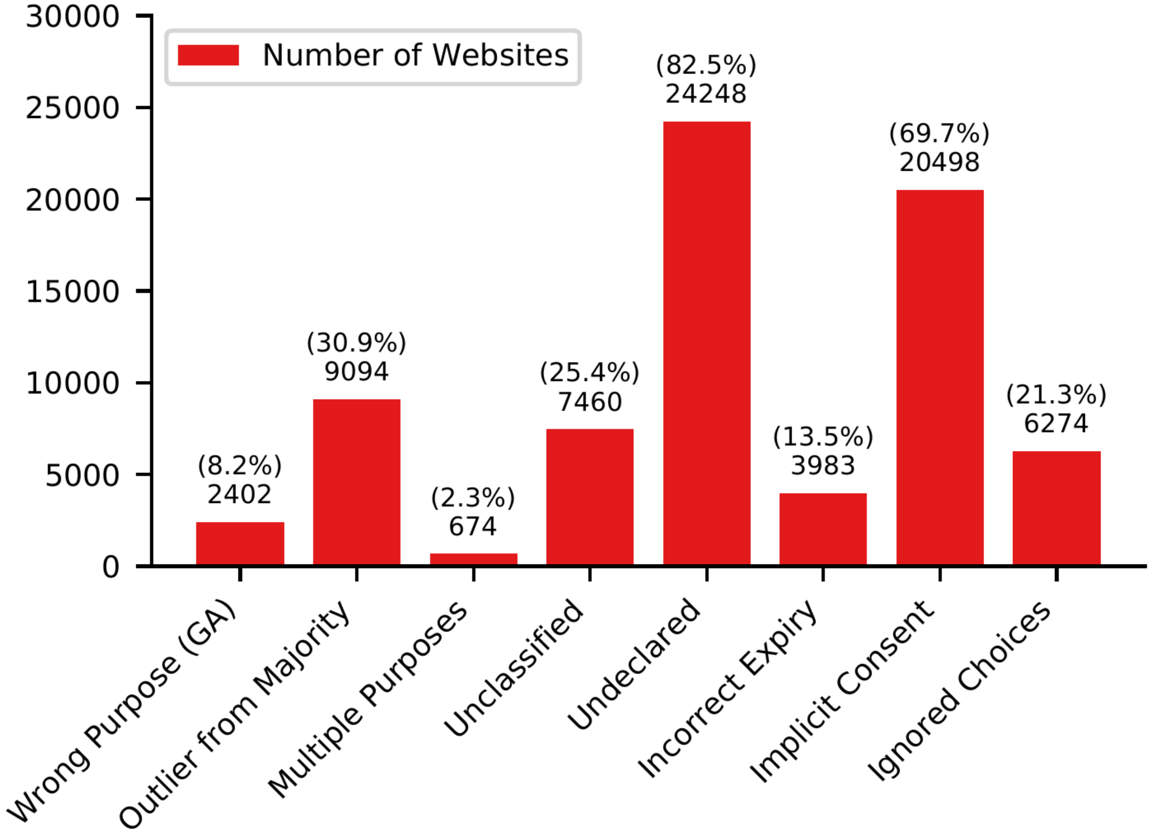
```
@inproceedings{bollinger2022automating,
  author = {Dino Bollinger and Karel Kubicek and Carlos Cotrini and David Basin},
  title = {Automating Cookie Consent and {GDPR} Violation Detection},
  booktitle = {31st USENIX Security Symposium (USENIX Security 22)},
  year = {2022},
  month = aug,
  pages = {2893--2910},
  isbn = {978-1-939133-31-1},
  publisher = {USENIX Association},
  url = {https://www.usenix.org/conference/usenixsecurity22/presentation/bollinger},
  address = {Boston, MA},
}
```

Cookie consent is fundamentally broken

Browser cookies are one of the most commonly used methods for tracking the session state of websites, and for tracking the identity of visitors. According to prior studies, between 80-90% of websites use cookies for user tracking, often without their knowledge. The EU government has attempted to address this issue through regulations mandating consent for data collection, in particular through the General Data Protection Regulation (*GDPR*) and the ePrivacy Directive.

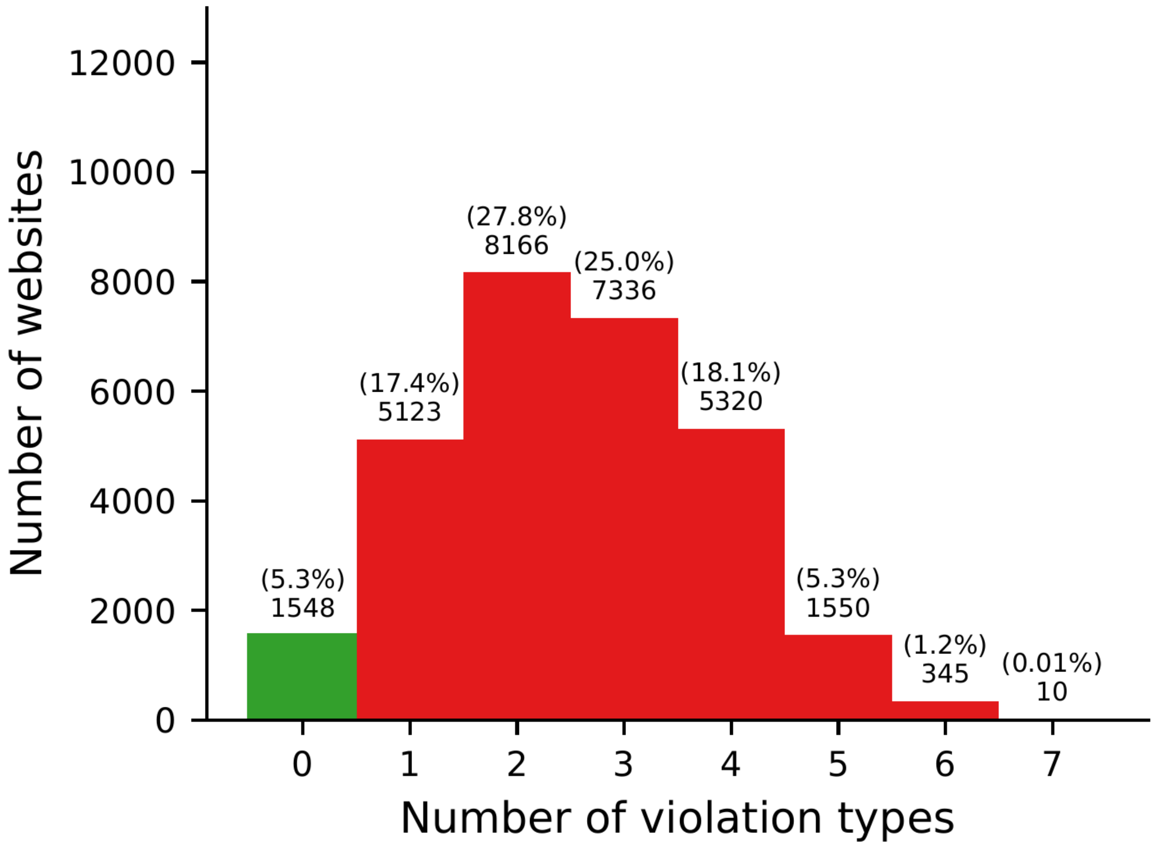
Despite these requirements, prior research has shown that less than half of all websites ask visitors for consent. Of the remaining websites, many violate even basic consent requirements. For instance, the majority of websites did not adhere to the opt-in requirement, and more than 10% stored affirmative consent before visitors could react to the consent notice. Some websites also stored affirmative consent despite explicit rejection, and many consent notices use dark patterns to nudge users into accepting all cookies.

In our analysis, we confirm the lack of GDPR compliance by extending and improving upon past research. We analyze the accuracy of the information displayed on cookie banners using a dataset collected from almost 30k websites. Specifically, we identify incorrect category assignments, misleading cookie expiration times, and assess the overall completeness of the consent mechanism. We define six novel methods to detect potential GDPR violations and extend two methods used in prior works. Of the selected domains, we find that 94.7% contained at least one potential violation. In 36.4%, we found at least one cookie with an incorrectly assigned purpose, and in 85.8%, there was at least one cookie with a missing declaration or missing purpose. 69.7% of sites assumed positive consent before it is given, and 21.3% created cookies despite negative consent. Our results indicate that consent notices are less compliant than previous research indicated.



The number of websites that show the respective type of violation. The first six are novel and have not been explored in prior work.

We support the legal claims by referring to relevant sections of GPDR and ePrivacy Directive and Planet49 case ruled by the EU Court of Justice.



This histogram shows the distribution of violation types per website, with the green bar representing the compliant ones. It does not include repetitions of a single type.

For the case of missing cookie declarations or purposes, we argue that the issues stem from neglect rather than malice. The cause is likely the lack of enforcement and web administrators who are not sufficiently familiar with the legal requirements. These violations can be addressed by providing regulatory authorities crawlers used in this study to improve enforcement of the GDPR. If you are a regulatory authority or any legal body interested in our work, please contact the authors.

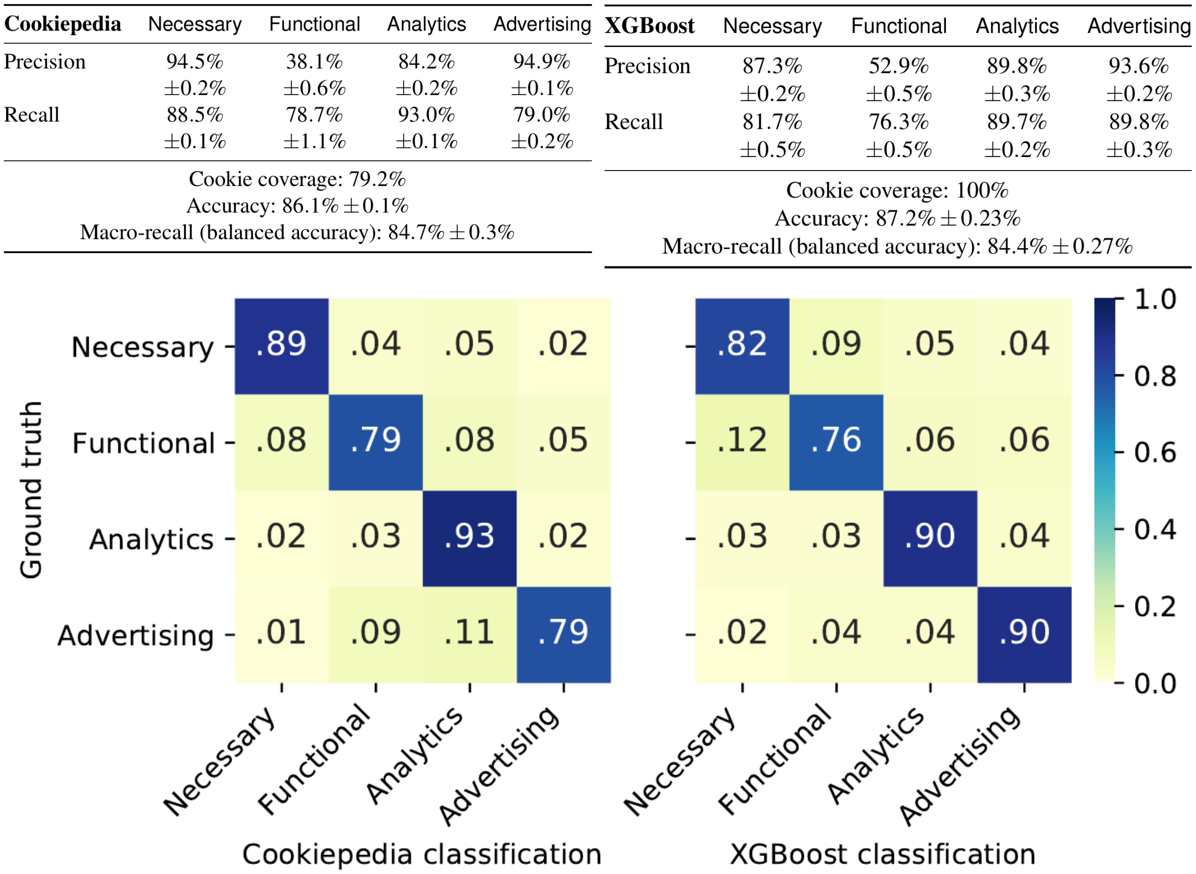
Client-side mitigation with extension CookieBlock

According to evidence from prior works and our own measurements, cookie consent practices are so often in violation of the GDPR that regulatory authorities cannot hope to keep up. Therefore, we provide users with a tool to enforce cookie consent on their web clients without requiring regulations to be enforced. We develop the browser extension CookieBlock, which classifies cookies by purpose, deleting those that the user rejects. In this way, the user can remove over 90% of all privacy-invasive cookies, without having to trust cookie banners. Previous attempts to provide users such control, like the P3P standard, failed due to a lack of willingness of website administrators to implement the functionality. We sidestep this problem by not relying on the cooperation of the websites at all.

In order to train the classifier model, we collected a dataset of approximately 304k cookies with purpose labels from 30k websites using selected Consent Management Platforms (CMP) to display the consent notice including cookie to purpose mapping. We retrieve the purposes the cookies are assigned to, and match these to the actual cookies that are created in the browser while visiting the website.

From the collected cookies, we extract statistically-rich, domain-specific features for each cookie. These features are determined from multiple attributes, including the name, domain, path, value, expiration date, as well as flags such as the “HttpOnly,” “Secure,” “SameSite,” and “HostOnly” properties.

For cookie classification, CookieBlock uses an ensemble of decision trees model, which is trained using the XGBoost library. We evaluate the model by comparing its performance to that of the Cookiepedia repository. Cookiepedia assigns purposes to cookies based on their name and was constructed manually for 10 years by human operators. We query this repository for purpose predictions and compare the results to the ground truth. In summary, we find that Cookiepedia achieves a balanced accuracy of 84.7%, while our XGBoost model achieves 84.4%. As such, our model is competitive with human expertise, showing that it is possible to automatically classify cookies by purpose using only the information available in the cookies themselves.



Cookiepedia	Necessary	Functional	Analytics	Advertising
Precision	94.5% ±0.2%	38.1% ±0.6%	84.2% ±0.2%	94.9% ±0.1%
Recall	88.5% ±0.1%	78.7% ±1.1%	93.0% ±0.1%	79.0% ±0.2%

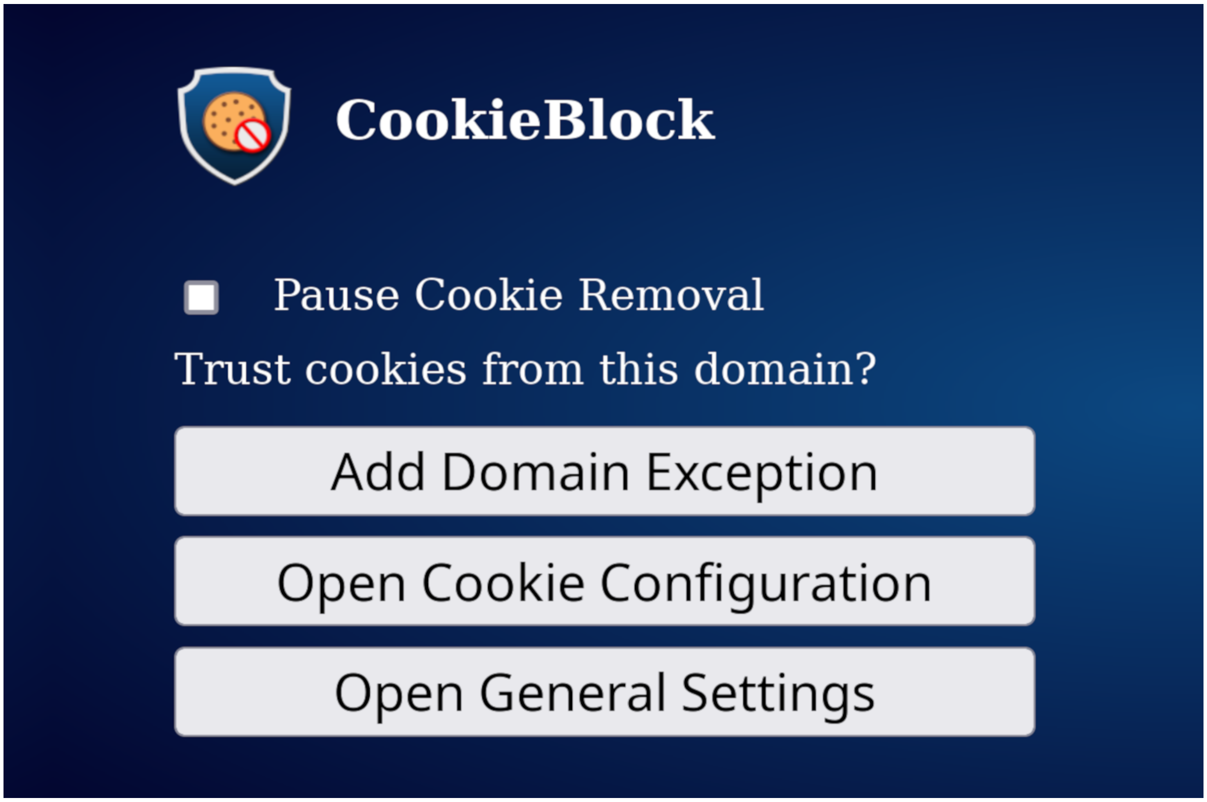
Cookie coverage: 79.2%
Accuracy: 86.1% ± 0.1%
Macro-recall (balanced accuracy): 84.7% ± 0.3%

XGBoost	Necessary	Functional	Analytics	Advertising
Precision	87.3% ±0.2%	52.9% ±0.5%	89.8% ±0.3%	93.6% ±0.2%
Recall	81.7% ±0.5%	76.3% ±0.5%	89.7% ±0.2%	89.8% ±0.3%

Cookie coverage: 100%
Accuracy: 87.2% ± 0.23%
Macro-recall (balanced accuracy): 84.4% ± 0.27%

Performance comparison of Cookiepedia to our automated XGBoost model, showing that our model is competitive with human expertise.

We evaluate CookieBlock on a set of 100 websites to quantify the impact the extension has on the browsing experience. CookieBlock causes no issues on 85% of the sites, minor problems involving non-essential website functions on 8%, and more substantial defects on 7%. The latter involve the user’s login status being lost due to the cookie removal. To resolve these problems, the user can selectively define website exemptions, and change the classification of cookies through CookieBlock’s interface.



CookieBlock interface consists of a simple popup and settings.

CookieBlock, the cookie purpose classification model, and the web crawler can help various parties:

- Web administrators can inspect the compliance of their website by detecting undeclared cookies and predicting purposes for currently unclassified cookies.
- Regulatory authorities can inspect the website’s compliance.
- Privacy- and web-researchers can use the model to classify the cookie types in follow-up studies.

Errata of paper (addressed in our version)

In Section 6.4, we claim *“This aligns with the results by Nouwens et al. [39], who found that 67.6% of 680 sites used implicit consent”* However, Nouwens et al. found the violations at **32.5%** of websites, therefore our results are not aligned with theirs. This is caused by entirely different sample of websites, where Nouwens et al. has a generic sample of websites while we focus only on websites with selected CMPs.

In section 4.1, we reported that Cookiepedia achieves 83.4% ballanced accuracy, while in the final measurement of our work, we observed 84.7%.

Q&A for users

- **Q:** Will there be Safari version?

  A: No, Safari lacks the necessary web extension APIs to intercept the creation of cookies in the browser. This is essential for gathering cookie data to predict a usage purpose, as well as to prevent undesired cookies from being stored in the browser. For those interested in technical details, they do not support any of these two methods for collecting cookie changes: cookies.onChanged and change event of cookie store API.

- **Q:** Is CookieBlock collecting any information from my browser? Does it store this in a remote host?

  A: No, the classification is done client-side. User can allow collection of a local cookie history, this is only used locally for improving the classifier accuracy. The data never leaves the browser's storage, and we adhere to the requirements of the GDPR. If you want to provide us feedback, please fill out the feedback form, or leave a review in the extension store or send us an email. If you observe a malfunctioning website, send us an email to cookie.block.extension@gmail.com, file an issue on GitHub, or make a pull-request with updated `known_cookies.json`.

- **Q:** CookieBlock breaks website xyz.com. What should I do?

  A: The easiest solution is to use the button "Add this site as an exception" in the CookieBlock popup. If the website does not start working immediately, you might need to clear cookies for the website. If even this does not help, you can also select "Pause cookie removal." No matter which technique you use, let us know about your issues by email to cookie.block.extension@gmail.com or an issue on GitHub.

- **Q:** What is the recommended settings for CookieBlock?

  A: The extension comes pre-installed with the recommended settings active. We recommend having Functionality cookies enabled to reduce potential website breakage. These types of cookies often enable the customization of websites without requiring to log in. Examples are, setting the language, enabling dark mode, etc. We also recommend to enable "Keep Track of Cookie History," which improves classification and makes the extension more efficient.

- **Q:** Does CookieBlock work outside of the EU?

  A: Yes, despite the fact that other countries do not have privacy regulations as advanced as the EU, the classification model works independently of your location. Note that CookieBlock will not cause websites to start displaying popups, only the cookies are being filtered.

  For websites other than in English, the model can have slightly reduced performance, as it cannot extract all meaningful features about the cookie content. However, the most useful features for the classification are language agnostic, so this might have only a minor impact. Enjoy what is called Brussels effect.

- **Q:** Can I use CookieBlock together with other extensions?

  A: Yes, we are not aware of any incompatibility with extensions such as uBlock Origin, AdBlock, Ghostery, Privacy Badger, Consent-O-Matic, or I don't care about cookies.

- **Q:** When and how does CookieBlock remove cookies?

  A: CookieBlock does not prevent websites to create cookies, as it needs to observe their content to classify them. The cookies are removed immediately after creation and classification. How long this takes depends when the browser informs us about the event of cookie creation, but usually it occurs after roughly 15 ms.

- **Q:** CookieBlock does not remove the cookie banners. How do I get rid of them?

A: We want to keep our extension as simple as possible and with only the purpose of removing the cookies. Recommended extensions that remove the popups are: Consent-O-Matic, I don’t care about cookies, or uBlock Origin with Annoyances filters (e.g., EasyList Cookie).

- **Q:** CookieBlock breaks Google services. What can I do?

  A: We try to make sure that Google services work as they are supposed to, but if you install CookieBlock to a profile with already some history, some issues may still arise. One solution is to sign out of your account and sign in again. If even that did not help, try removing cookies of your browser, a fresh profile should work correctly.

Q&A for developers

- **Q:** CookieBlock breaks my website. How can I make my website compatible with CookieBlock?

  A: You can try CookieBlock on your website and check whether the classification of cookies matches your expectations. Consider whether you are not overusing cookies to multiple purposes. If not, you can report the problems, or directly contribute to CookeiBlock (pull request on GitHub) with the exceptions.

- **Q:** My website is in the published dataset, and according to it, it contains violations. Can you redact it?

  A: While some of the violation detection methods can produce false positives, given the vast non-compliance, it is likely that your website does violate some of the consent requirements. We also appreciate your effort, given that your website at least uses a CMP giving user choices, which is better than 96% of the Internet. You can contact the CMP or the web developer to inspect what can be improved. Anyway, you can contact us at cookie.block.extension@gmail.com, and we can figure out the removal from the dataset.

Q&A for researchers/others

- **Q:** How severe are the detected violations?

  A: While we refer to the violations as *potential*, we do so as only a judicial ruling can provide the legal certainty as to whether they are true legal violations. All of the violations imply that the user consents to incorrect information, and it requires an individual inspection to declare them as neglect or malice. However, severe violations as marking Google Analytics cookies as necessary or unclassified or even omitting to declare them forces users to accept them. In such a case, the website might do even worse than the website without a cookie banner, as it gives users a false impression of choice.

Media

- NZZ am Sonntag: German, paywall, contact us for full text
- ETH news: English and German version.

Acknowledgement

The authors would like to thank:

- Midas Nouwens for discussion about the initial idea of the paper.
- Reviewers for their feedback.

Updates

- *August 7, 2023:* Errata for mistakes found in the paper.
- *July 26, 2023:* Recording link.
- *August 17, 2022:* We won USENIX’s distinguish artifact reward.
- *June 8, 2022:* FAQ clarifications, artifacts link update, images improved.
- *March 23, 2022:* More FAQ and media references.
- *February 6, 2022:* Added slides from invited talk at EPFL.
- *November 17, 2021:* We received all 3 artifact badges: Artifact Available, Artifact Functional, and Artifact Reproduced.
- *November 11, 2021:* Paper released by USENIX.
- *October 15, 2021:* Camera-ready paper version.
- *September 21, 2021:* The initial version of this page.