Comparing Privacy Policies of Government Agencies and Companies: A Study using Machine-learning-based Privacy Policy Analysis Tools

Razieh Nokhbeh Zaeem[a] and K. Suzanne Barber[b]

Center for Identity at the University of Texas at Austin, Austin, TX, U.S.A.

Keywords: Privacy Policy, Privacy Enhancing Technologies, Machine Learning, Government Agencies, Companies, European Union, United States.

Abstract: Companies and government agencies are motivated by different missions when collecting and using Personally Identifiable Information (PII). Companies have strong incentives to monetize such information, whereas government agencies are generally not-for-profit. Besides this difference in missions, they are subject to distinct regulations that govern their collection and use of PII. Yet, do privacy policies of companies and government agencies reflect these differences and distinctions? In this paper, we take advantage of two of the most recent machine-learning-based privacy policy analysis tools, Polisis and PrivacyCheck, and five corpora of over 800 privacy policies to answer this question. We discover that government agencies are considerably better in protecting (or not collecting for that matter) sensitive financial information, Social Security Numbers, and user location. On the other hand, many of them fail to directly address children's privacy or describe security measures taken to protect user data. Furthermore, we observe that E.U government agencies perform well, with respect to notifying users of policy change, giving users the right to edit/delete their data, and limiting data retention. Our work confirms the common-sense understanding that government agencies collect less personal information than companies, but discovers nuances, as listed above, along the way. Finally, we make our data publicly available, enhancing reproducibility and enabling future analyses.

1 INTRODUCTION

Privacy policies have become the de facto way of communicating privacy practices of companies, government agencies, or any other organization with their consumers/clients. These privacy policies outline how the company or government agency handles, shares, discloses, and uses Personally Identifiable Information (PII) of its consumers or clients. PII is defined as "any information relating to an identified or identifiable natural person"[1] (Union,) such as name, email address, and credit card number.

Companies and government agencies alike collect PII. They are, however, motivated by different missions and governed by different regulations with regard to PII. The research question we answer in this paper is that *if* (and *how*) companies and government agencies are different in their *privacy policies*.

1.1 Missions of Companies and Government Agencies

Companies are in the business of making money. As a result, their collection and use of PII involves monetization, for example, through advertisement. On the other hand, government agencies are generally not for-profit entities and therefore may have much less motivation to collect personal data and sell or otherwise monetize it. Therefore, privacy policies of companies and government agencies, one might argue, must be different. This difference in missions, however, does not negate the importance of our research question. We aim to exactly quantify such differences through statistically significant results and find unexpected similarities and differences as well.

1.2 Regulations Governing Privacy Policies

Many general data privacy regulations apply only to companies:

[a] https://orcid.org/0000-0002-0415-5814

[b] https://orcid.org/0000-0003-2906-6583

[1] https://gdpr-info.eu

Zaeem, R. and Barber, K.
Comparing Privacy Policies of Government Agencies and Companies: A Study using Machine-learning-based Privacy Policy Analysis Tools.
DOI: 10.5220/0010180700290040
In *Proceedings of the 13th International Conference on Agents and Artificial Intelligence (ICAART 2021) - Volume 2*, pages 29-40
ISBN: 978-989-758-484-8

- The General Data Protection Regulation (GDPR) is the newest regulation in the E.U. law on data protection and privacy. The key principles of the GDPR are (1) Lawfulness, fairness, and transparency, (2) Purpose limitation, (3) Data minimization, (4) Accuracy, (5) Storage limitation, (6) Security, and (7) Accountability.
- The California Consumer Privacy Act (CCPA) is a state statute to enhance privacy rights and consumer protection for residents of California.
- The Federal Trade Commission (FTC) Fair Information Practice Principles (FIPP) are recommendations, though not legally enforced, for maintaining privacy-friendly, consumer-oriented data collection practices and include Notice, Choice, Access, and Integrity.
- The Children's Online Privacy Protection Act (COPPA) protects the personal information of children under 13.
- The Gramm-Leach-Bliley Act (GLBA) requires financial institutions (i.e. companies) to explain their information-sharing practices to their customers and to safeguard sensitive data.
- The Health Insurance Portability and Accountability Act (HIPPA) applies to health care providers, suppliers and vendors (business associates).
- The Telephone Consumer Protection Act (TCPA) regulates the collection of information by telephone service providers.

Some regulations, however, are dedicated to ensuring the privacy of data when collected by the government. For example, in the U.S.:

- The Privacy Act of 1974 is the primary law in the U.S. that governs government collection, maintenance, use, and dissemination of PII by federal agencies.
- The Freedom of Information Act (FOIA) governs the collection, maintenance, use, and dissemination of PII that is maintained in systems of records by federal agencies.
- Federal Information Security Management Act (FISMA) mandates that each federal agency implements an information security program for the information and information systems that support the operations and assets of the agency.
- The Electronic Communications Privacy Act (ECPA) restricts the government's access and disclosure of electronic communication.

In this paper, we investigate whether contrasting missions and different regulations have resulted in differing privacy policies among government agencies versus corporations. We study a total of 550 privacy policies of companies and 285 policies of government agencies through privacy policy analysis tools.

1.3 Automatic Analysis of Privacy Policies

Privacy policy analysis tools utilize machine learning and natural language processing to automatically extract information from privacy policies. Because privacy policies are long, cumbersome to read, and difficult to comprehend for their final consumers (Ermakova et al., 2014; Milne et al., 2006; Graber et al., 2002; McDonald and Cranor, 2008), researchers have developed tools to summarize these policies automatically. Some of the most recent such tools are Polisis[2] (Harkous et al., 2018) and PrivacyCheck[3] (Zaeem et al., 2018; Zaeem and Barber, 2017).

Interestingly, these tools have found a new use in addition to assisting final consumers in understanding privacy policies: researchers have leveraged these tools to automatically analyze huge corpora of privacy policies and study their statistics. For instance, both Polisis and PrivacyCheck have been utilized to study the effect of the GDPR on the privacy landscape (Linden et al., 2018; Zaeem and Barber, 2020).

1.4 Comparing Privacy Policies of Government Agencies and Companies

In this work, we leverage privacy policy analysis tools to automatically compare privacy policies of companies and government agencies. We obtained, from independent research groups, two privacy policy analysis tools, and three corpora (totaling 550) of corporate privacy policies. In addition, we crawled the official comprehensive lists of the United States federal government agencies and the European Union agencies, and hence added two respective corpora (totaling 285) of government agencies' privacy policies.

We find that both U.S. and E.U. agencies protect PII (such as Credit Card Number and Location data) better than companies or even do not collect them in the first place. On the other hand, many of them fail to directly address children's privacy or describe security measures taken to protect user data. We also

[2] Available online at https://pribot.org/polisis.

[3] Available online at https://identity.utexas.edu/privacycheck-for-google-chrome.

find that E.U. government agencies' polices are superior to their U.S. counterparts, with respect to notifying users of policy changes, giving users the right to edit/delete their data, and limiting data retention—all of which are GDPR tenets.

We make the following contributions:

1. We are the first to compare privacy policies of government agencies with companies, whether manually or automatically.
2. We utilize inferential statistics to measure the significance and degree of association in our results.
3. We made publicly available our collected and data-mined source data, including links to all privacy policies, downloaded text used in the experiments, and the analytical results of running privacy policy analysis tools on them.

The rest of this paper is organized as follows. Sections 2 elaborates on the privacy policy analysis tools. Section 3 discusses the privacy policy corpora employed in this analysis. Section 4 presents the experiments and results. Section 6 covers the related work, and finally, Section 7 concludes the paper.

2 PRIVACY POLICY ANALYSIS TOOLS

We seek to compare privacy policies in an automated and scalable manner through the use of privacy policy analysis tools. Therefore, we utilize the most recent tools we found in the literature that automatically summarize privacy policies:

1. PrivacyCheck (Zaeem et al., 2018; Zaeem and Barber, 2017) was developed at the University of Texas at Austin. PrivacyCheck is a machine learning tool, released as web browser extensions for Google Chrome and Mozilla Firefox, that automatically summarizes a privacy policy to answer ten fundamental questions concerning an organization's PII security and privacy protections. The Chrome PrivacyCheck extension currently has 901 users.
2. Polisis (Harkous et al., 2018) is a browser extension, also available for Google Chrome and Mozilla Firefox, that takes advantage of deep learning to summarize what PII the privacy policy claims to be collecting and sharing. The Chrome Polisis extension currently has 1,011 users.

2.1 Background: PrivacyCheck

PrivacyCheck (Zaeem et al., 2018; Zaeem and Barber, 2017) utilizes classification methods, particularly Naive Bayes, to answer ten basic questions about the privacy and security of user data according to a privacy policy. PrivacyCheck was trained using 400 training policies against manually extracted answers for these ten questions. When analyzing a new privacy policy, PrivacyCheck employs trained classifiers to assign an answer to each of the ten questions.

These ten questions were compiled from previous work, e.g., the work of the Organization for Economic Co-operation and Development (Regard, 1980), and the FTC FIPP (FTC, 2000). According to the PrivacyCheck extension for the Google Chrome web browser, these ten questions are as follows.

1. Email Address: how does the site handle your email address?
2. Credit Card Number: how does the site handle your credit card number and home address?
3. Social Security Number: how does the site handle your Social Security number?
4. Ads and Marketing: does the site use or share your personally identifiable information for marketing purposes?
5. Location: does the site track or share your location?
6. Collecting PII of Children: does the site collect personally identifiable information from children under 13?
7. Sharing with Law Enforcement: does the site share your information with law enforcement?
8. Policy Change: does the site notify you or allow you to opt out when their privacy policy changes?
9. Control of Data: does the site allow you to edit or delete your information from its records?
10. Data Aggregation: does the site collect or share aggregated data related to your identity or behavior?

The possible answer to each of the above questions is one of the three risk levels: Green, Yellow, and Red, for low, medium, and high risk, respectively. The risk levels for each question, according to the same Chrome extension, are as displayed in Table 1 (from previous work (Zaeem et al., 2018)).

2.2 Background: Polisis

At its core, Polisis (Harkous et al., 2018) is a neural network classifier trained on 130,000 privacy policies retrieved from the Google Play store. Polisis

Table 1: Risk levels for privacy factors, from the authors of PrivacyCheck (Zaeem et al., 2018).

Privacy Factor	Green Risk Level	Yellow Risk Level	Red Risk Level
1. Email Address	Not asked for	Used for the intended service	Shared w/ third parties
2. Credit Card Number	Not asked for	Used for the intended service	Shared w/ third parties
3. Social Security Number	Not asked for	Used for the intended service	Shared w/ third parties
4. Ads and Marketing	PII not used for marketing	PII used for marketing	PII shared for marketing
5. Location	Not tracked	Used for the intended service	Shared w/ third parties
6. Collecting PII of Children	Not collected	Not mentioned	Collected
7. Sharing w/ Law Enforcement	PII not recorded	Legal docs required	Legal docs not required
8. Policy Change	Posted w/ opt out option	Posted w/o opt out option	Not posted
9. Control of Data	Edit/delete	Edit only	No edit/delete
10. Data Aggregation	Not aggregated	Aggregated w/o PII	Aggregated w/ PII

segments a privacy policy and automatically annotates each segment with a set of labels, classifying segments based on coarse- and fine-grained classifications.

The questions Polisis answers are based on the ten privacy taxonomy of Wilson et al. (OPP-115) (Wilson et al., 2016a) and are as follows:

1. First Party Collection/Use: how and why does the site collects PII?
2. Third Party Sharing/Collection: how is PII shared with or collected by third parties through this site?
3. User Choice/Control: what are the choices and control options available to users?
4. User Access, Edit, & Deletion: if and how do users may access, edit, or delete their information?
5. Data Retention: how long is user information stored?
6. Data Security: how is user information protected?
7. Policy Change: if and how will users be informed about changes to the privacy policy?
8. Do Not Track: are Do Not Track signals for online tracking and advertising honored?
9. International & Specific Audiences: practices that pertain only to a specific group of users (e.g., children, Europeans).
10. Other: additional sub-labels not covered above.

Each of the above top-level taxonomies are further broken down to lower level sets of policy attributes with a possible set of values or answers. The list of answers is extensive and can be found elsewhere (Wilson et al., 2016a).

3 CORPORA

In this section, we first review the three corpora of corporate privacy policies, followed by the two corpora of government agencies' privacy policies.

3.1 Corpora of Companies

We experiment with three corpora of online privacy policies of 400, 50, and 100 companies.

1. **The Stock Market Companies.** The authors of PrivacyCheck complied the first corpus of 400 policies by considering 10% of all the companies listed on NYSE, Nasdaq, and AMEX stock markets (Zaeem et al., 2018; Zaeem and Barber, 2017).
2. **The Web Search Companies.** The authors of PrivacyCheck selected the second corpus of 50 privacy policies through a web search (Zaeem et al., 2018; Zaeem and Barber, 2017) for "privacy policy".
3. **The Mobile App Companies.** The authors of Polisis and Pribots crawled the Google Play Store for privacy policies of mobile applications. We obtained our third corpus from them, and after accounting for duplicate and similar policy texts, we used a corpus of 100 unique policies (Harkous et al., 2018; Harkous et al., 2016; Zaeem and Barber, 2020) for this research effort.

3.2 Corpora of Government Agencies

We gathered two corpora of online privacy policies of 249 U.S. and 36 E.U. government agencies.

1. Starting from the official comprehensive list of federal U.S. government agencies at https://www.usa.gov/federal-agencies, we crawled the web to list one URL of a privacy policy per each U.S.

government agency. After deleting repetitive URLs (as some agencies link to the same privacy policy), we distilled a corpus of 249 privacy policies. We further manually verified that all the collected links do indeed point to a privacy policy.

2. Starting from the official listing of E.U. agencies at https://europa.eu/european-union/about-eu/agencies_en, we reached all the 51 E.U. agencies' websites and fetched all of their privacy policies, also commonly known as "personal data protection" or "legal notice" on these websites. After deleting duplicates, 43 policies remained including 7 pdf files. However, neither Polisis nor PrivacyCheck can analyze pdf files. Therefore, we ignored the pdf files and used the remaining 36 privacy policies of E.U. agencies.

4 RESULTS

With help from their corresponding authors to access their API, we ran Polisis and PrivacyCheck on the five corpora in late 2019 and early 2020 and recorded the results.

4.1 PrivacyCheck Results

Figures 1 and 2 show how PrivacyCheck scores the U.S. and E.U. government agencies' privacy policies, respectively, while Figures 3, 4, and 5 display the result of running PrivacyCheck on corporate privacy policies in the three corpora. Each bar displays the number of privacy policies in the corresponding corpus with a given risk level color. The figures are scaled to be visually comparable in terms of the percentage in each corpus.

We utilize statistical analysis to study the differences between the corpora. Using cross tabulation in the IBM SPSS software[4], we measure the statistical significance of the PrivacyCheck returned risk levels (Tables 2 to 11). **Throughout this paper:** we use $\alpha = .01$. For all of the PrivacyCheck factors $p < .001$, so the results are significant at $p < .01$.

We further measure Cramer's V, one of the most common chi-square-based measures of nominal association. V ranges between 0 and 1, with 0 indicating no association and 1 showing complete association. The results of the chi-square test for statistical significance and V for association measurement are shown in the last row of each table.

With respect to the following privacy factors, both U.S. and E.U. government agencies protect PII better than companies:

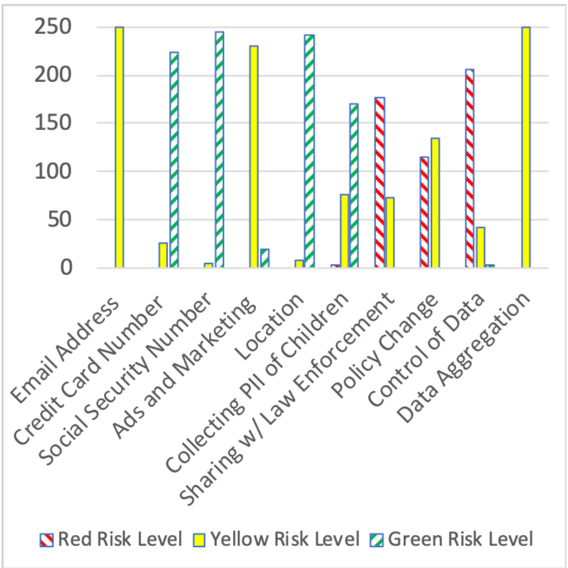

Figure 1: Privacy policies of the 249 U.S. government agencies: the distribution of PrivacyCheck risk levels.

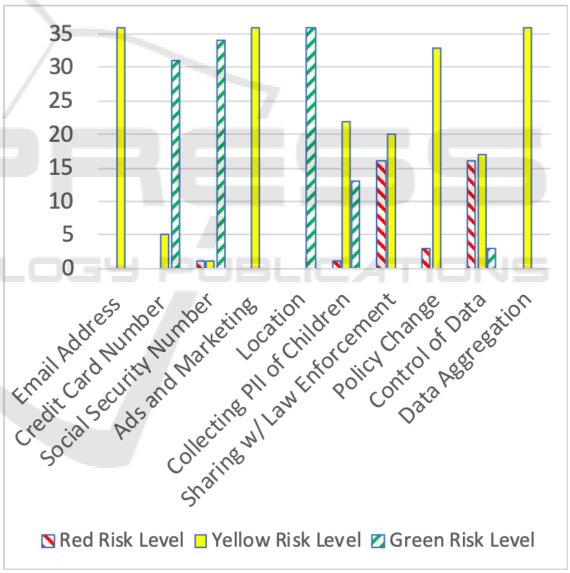

Figure 2: Privacy policies of the 36 E.U. government agencies: the distribution of PrivacyCheck risk levels.

Credit Card Number. (Table 3): There is no policy with a red risk level (PII shared with third parties) in either of the government corpora. In addition, many more are at the green level (PII not asked for), compared to the corporate policies.

Social Security Number. (Table 4): Almost all of the government agencies' policies are at the green level—they do not collect this PII.

Location. (Table 6): The vast majority of government agencies do not track location and are at the

[4] https://www.ibm.com/analytics/spss-statistics-software

Table 2: PrivacyCheck results for Email Address.

	Red	Yellow	Green
EU	0	36	0
% within Corpus	0.00%	100.00%	0.00%
US	0	249	0
% within Corpus	0.00%	100.00%	0.00%
APP	4	96	0
% within Corpus	4.00%	96.00%	0.00%
STOCK	55	335	10
% within Corpus	13.80%	83.80%	2.50%
WEB	2	48	0
% within Corpus	4.00%	96.00%	0.00%
Total	61	764	10
% within Corpus	7.30%	91.50%	1.20%
$\chi^2(8, N = 835) = 61.89, p < .001, V = .193$.			

Table 3: PrivacyCheck results for Credit Card Number.

	Red	Yellow	Green
EU	0	5	31
% within Corpus	0.00%	13.90%	86.10%
US	0	26	223
% within Corpus	0.00%	10.40%	89.60%
APP	2	53	45
% within Corpus	2.00%	53.00%	45.00%
STOCK	32	151	217
% within Corpus	8.00%	37.80%	54.30%
WEB	2	34	14
% within Corpus	4.00%	68.00%	28.00%
Total	36	269	530
% within Corpus	4.30%	32.20%	63.50%
$\chi^2(8, N = 835) = 153.89, p < .001, V = .304$.			

Table 4: PrivacyCheck results for Social Security Number.

	Red	Yellow	Green
EU	1	1	34
% within Corpus	2.80%	2.80%	94.40%
US	0	4	245
% within Corpus	0.00%	1.60%	98.40%
APP	3	6	91
% within Corpus	3.00%	6.00%	91.00%
STOCK	42	69	289
% within Corpus	10.50%	17.30%	72.30%
WEB	4	1	45
% within Corpus	8.00%	2.00%	90.00%
Total	50	81	704
% within Corpus	6.00%	9.70%	84.30%
$\chi^2(8, N = 835) = 91.44, p < .001, V = .234$.			

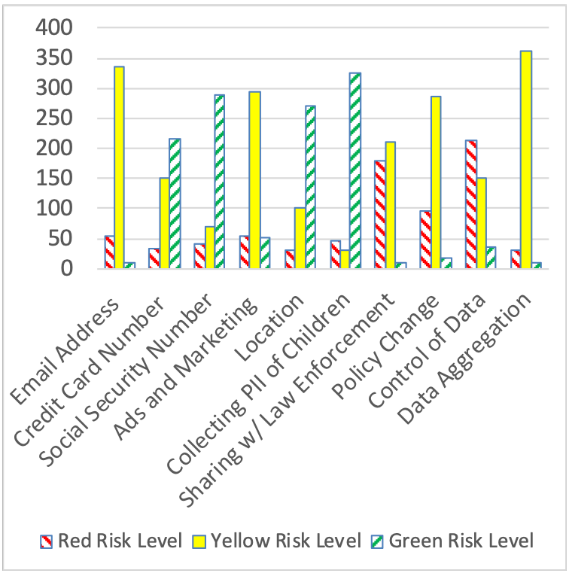

Figure 3: Privacy policies of the 400 stock market companies: the distribution of PrivacyCheck risk levels.

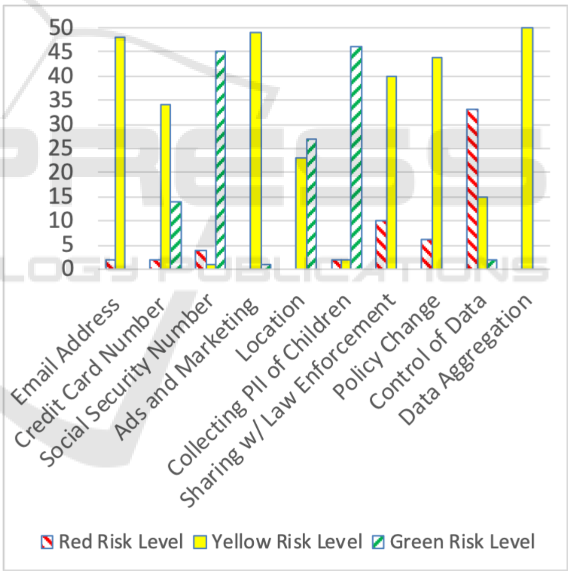

Figure 4: Privacy policies of the 50 web search companies: the distribution of PrivacyCheck risk levels.

green level.

With respect to one factor, government agencies perform poorly in comparison. **Collecting PII of Children** (Table 7): Above 80% of the three corpora of corporate policies are rated at the green level—they do not collect children's information without parental consent. However, the government agencies have a lot of policies at the yellow level—i.e., no mention of children's privacy. The underlying reason might be that government agencies are not presumed to target children as users.

Table 5: PrivacyCheck results for Ads and Marketing.

	Red	Yellow	Green
EU	0	36	0
% within Corpus	0.00%	100.00%	0.00%
US	0	230	19
% within Corpus	0.00%	92.40%	7.60%
APP	1	99	0
% within Corpus	1.00%	99.00%	0.00%
STOCK	55	293	52
% within Corpus	13.80%	73.30%	13.00%
WEB	0	49	1
% within Corpus	0.00%	98.00%	2.00%
Total	56	707	72
% within Corpus	6.70%	84.70%	8.60%
$\chi^2(8, N = 835) = 92.73, p < .001, V = .236.$			

Table 6: PrivacyCheck results for Location.

	Red	Yellow	Green
EU	0	0	36
% within Corpus	0.00%	0.00%	100.00%
US	0	7	242
% within Corpus	0.00%	2.80%	97.20%
APP	1	28	71
% within Corpus	1.00%	28.00%	71.00%
STOCK	30	100	270
% within Corpus	7.50%	25.00%	67.50%
WEB	0	23	27
% within Corpus	0.00%	46.00%	54.00%
Total	31	158	646
% within Corpus	3.70%	18.90%	77.40%
$\chi^2(8, N = 835) = 126.50, p < .001, V = .275.$			

Table 7: PrivacyCheck results for Collecting PII of Children.

	Red	Yellow	Green
EU	1	22	13
% within Corpus	2.80%	61.10%	36.10%
US	3	76	170
% within Corpus	1.20%	30.50%	68.30%
APP	4	4	92
% within Corpus	4.00%	4.00%	92.00%
STOCK	45	30	325
% within Corpus	11.30%	7.50%	81.30%
WEB	2	2	46
% within Corpus	4.00%	4.00%	92.00%
Total	55	134	646
% within Corpus	6.60%	16.00%	77.40%
$\chi^2(8, N = 835) = 151.83, p < .001, V = .302.$			

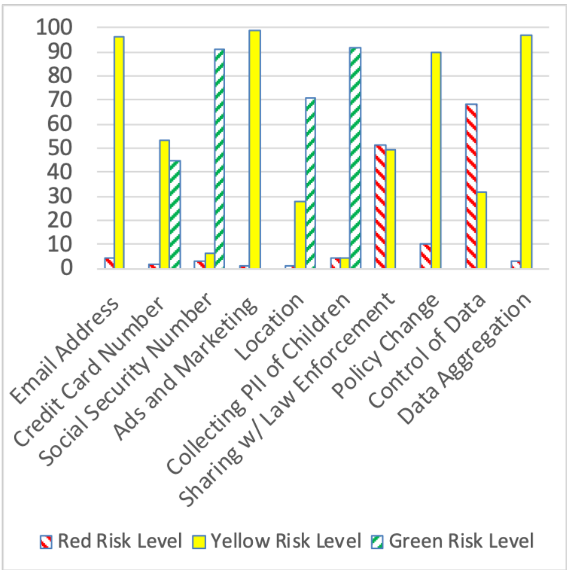

Figure 5: Privacy policies of the 100 mobile app companies: the distribution of PrivacyCheck risk levels.

Table 8: PrivacyCheck results for Sharing with Law Enforcement.

	Red	Yellow	Green
EU	16	20	0
% within Corpus	44.40%	55.60%	0.00%
US	176	73	0
% within Corpus	70.70%	29.30%	0.00%
APP	51	49	0
% within Corpus	51.00%	49.00%	0.00%
STOCK	179	210	11
% within Corpus	44.80%	52.50%	2.80%
WEB	10	40	0
% within Corpus	20.00%	80.00%	0.00%
Total	432	392	11
% within Corpus	51.70%	46.90%	1.30%
$\chi^2(8, N = 835) = 74.53, p < .001, V = .211.$			

Policy Change. (Table 9): Notably, E.U. agencies and corporate corpora all do better than U.S. government agencies with respect to this factor. Many U.S. agencies are rated at the red level as they might change their policies without notice.

Finally, for **Ads and Marketing** (Table 5) the majority of the policies are ranked at the yellow level which means PII is used to communicate/advertise their own services. The STOCK corpus is notably different, with policies at both green (PII not used for marketing) and red levels (PII shared for marking). The U.S. government agencies also have some policies at the green level.

Table 9: PrivacyCheck results for Policy Change.

	Red	Yellow	Green
EU	3	33	0
% within Corpus	8.30%	91.70%	0.00%
US	115	134	0
% within Corpus	46.20%	53.80%	0.00%
APP	10	90	0
% within Corpus	10.00%	90.00%	0.00%
STOCK	97	285	18
% within Corpus	24.30%	71.30%	4.50%
WEB	6	44	0
% within Corpus	12.00%	88.00%	0.00%
Total	231	586	18
% within Corpus	27.70%	70.20%	2.20%

$\chi^2(8, N = 835) = 92.51, p < .001, V = .235.$

Table 10: PrivacyCheck results for Control of Data.

	Red	Yellow	Green
EU	16	17	3
% within Corpus	44.40%	47.20%	8.30%
US	206	41	2
% within Corpus	82.70%	16.50%	0.80%
APP	68	32	0
% within Corpus	68.00%	32.00%	0.00%
STOCK	212	151	37
% within Corpus	53.00%	37.80%	9.30%
WEB	33	15	2
% within Corpus	66.00%	30.00%	4.00%
Total	535	256	44
% within Corpus	64.10%	30.70%	5.30%

$\chi^2(8, N = 835) = 77.34, p < .001, V = .215.$

Table 11: PrivacyCheck results for Data Aggregation.

	Red	Yellow	Green
EU	0	36	0
% within Corpus	0.00%	100.00%	0.00%
US	0	249	0
% within Corpus	0.00%	100.00%	0.00%
APP	3	97	0
% within Corpus	3.00%	97.00%	0.00%
STOCK	30	361	9
% within Corpus	7.50%	90.30%	2.30%
WEB	0	50	0
% within Corpus	0.00%	100.00%	0.00%
Total	33	793	9
% within Corpus	4.00%	95.00%	1.10%

$\chi^2(8, N = 835) = 37.87, p < .001, V = .151.$

We refrain from making conclusions about the factors with the lowest values of V, as we also did not observe huge differences between government agencies and companies: **Email Address** (Table 2, mostly yellow which means PII gathered and used for the intended service), **Sharing with Law Enforcement** (Table 8, yellow/red which mean PII shared with law enforcement with/without legal documents), **Control of Data** (Table 10, mostly red and yellow which mean no deletion of data is permitted but editing maybe allowed, with E.U. government agencies performing the best and U.S. government agencies the worst), and **Data Aggregation** (Table 11, yellow which means PII is aggregated).

In sum, both U.S. and E.U. agencies protect Credit Card Number, Social Security Number, and Location better than companies, and often fail to mention how they handle PII of children. E.U. government agencies' polices are better than their U.S. counterparts with respect to notifying the user after the policy changes.

4.2 Polisis Results

Polisis, when ran in its browser extension form, returns assessments about a privacy policy. We list all the assessments Polisis returned on privacy policies of our corpora in Table 12. In this table, we took advantage of the annotation scheme of the OPP-115 Corpus[5] (Wilson et al., 2016a), the underlying dataset of Polisis, to group these assessments based on the categories of Polisis referenced in Section 2.2. When we cross-tabulate whether an assessment was returned for a policy in the corpora and run the chi-square test, the results are not statistically significant for some of the assessments. Table 12 also shows whether statistical significance was achieved and the value of Cramer's V when statistically significant.

Figure 6 displays what percentage of each corpus received an assessment from Polisis. Assessments are grouped together based on their categories. We summarize our findings, only for those assessments that achieved statistical significance, as follows:

1. First Party Collection/Use. As evident from Figure 6, companies consistently collect *several types of PII* while government agencies do not.

2. Third Party Sharing/Collection. All the three corporate corpora show higher percentage of *several types of PII shared* and *PII shared for marketing*, even though they also have higher percentage of *some PII aggregated before sharing*.

[5] https://usableprivacy.org/data

Table 12: Assessment index for Polisis. We do not report V where statistical significance is not achieved.

Short name	Assessment	p	Significant?	V
	1. First Party Collection/Use			
No PII collected	No personally identifiable information is collected.	.012	×	—
PII collected for ads	Personal information is collected for marketing reasons.	.101	×	—
PII only collected for service	Personal information is only collected for providing the service itself.	.709	×	—
Tracks users on other websites	The service tracks users on other websites.	.013	×	—
Several types of PII collected	Several types of personal information types can be collected.	$< .001$	✓	.288
	2. Third Party Sharing/Collection			
No PII shared with third parties	The policy states that third parties do not receive personal information.	.391	×	—
PII only shared for service	Personal information is only shared with third parties for providing the service.	.666	×	—
Several types of PII shared	Several types of personal information types are shared with third parties.	.004	✓	.137
Ad providers track users	The policy allows ad providers and analytics firms to track users on the site.	.520	×	—
Only aggregated PII shared	Third parties only receive aggregated or anonymized information.	.568	×	—
PII shared for marketing	Personal information may be shared with third parties for marketing reasons.	.001	✓	.182
Health data shared	Health data might be shared with third parties.	.674	×	—
Some PII aggregated before sharing	Some data is anonymized or aggregated before sharing with third parties.	$< .001$	✓	.181
No sharing in certain conditions	In certain conditions, data is not shared.	.569	×	—
Location or address shared	Location or address data may be shared with third parties.	.252	×	—
Third parties do not receive PII	Third parties do not receive personally identifiable information.	.647	×	—
	3. User Choice/Control			
No choice	The only choices in the policy are not to use the service.	.646	×	—
Provides opt-in choices	The policy provides opt-in choices.	.001	✓	.156
Provides opt-out choices	The policy provides opt-out choices.	.001	✓	.271
Clear links to control data	The policy offers you clear links to control your data.	$< .001$	✓	.198
	4. User Access, Edit, & Deletion			
Access and deletion of PII	You can request access and deletion of personal data.	$< .001$	✓	.392
	5. Data Retention			
PII retained for defined period	Some data is retained for a well-defined period.	.007	✓	.133
PII retained indefinitely	Some data might be retained indefinitely.	$< .001$	✓	.216
	6. Data Security			
Covers security measures	The policy covers security measures in details.	.002	✓	.144
	7. Policy Change			
Notice when policy changes	There will be a clear notice when the policy changes.	.447	×	—
	8. Do Not Track			
None				
	9. International & Specific Audiences			
Section on children's privacy	The policy has a special section on respecting children's privacy.	$< .001$	✓	.249
	10. Other			
None				

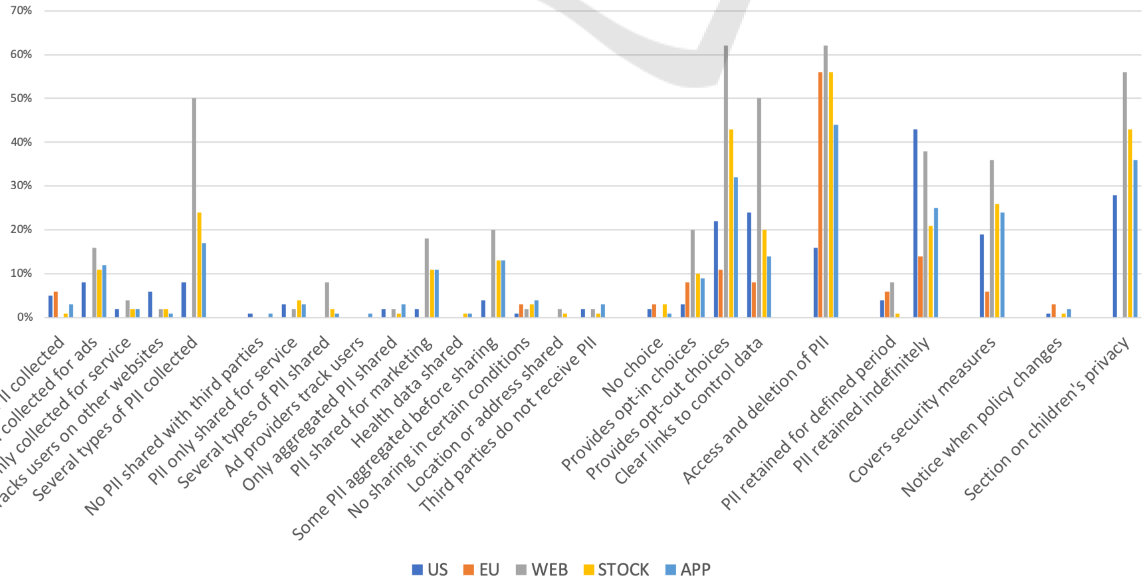

Figure 6: The percentage of each corpus that received an assessment from Polisis (Table 12).

3. User Choice/Control. Polisis generated mixed assessments about user choice and control in our corpora. There is no clear winner among these privacy policy corpora when taking into account all the choice/control assessments.

4. User Access, Edit, & Deletion. U.S. government agencies perform poorly in comparison with their E.U. counterparts and companies when providing *access to edit/delete PII*.

5. Data Retention. E.U. government agencies have the lowest percentage of *retaining PII indefinitely*.

6. Data Security. E.U. government agencies, however, often fail to outline exact *security measures* taken to protect user data. U.S. agencies are also trailing behind companies.

7. Policy Change. We observe no meaningful difference among our corpora as the results are not significant. The PrivacyCheck results, however, showed that U.S. government agencies sometimes fail to notify users of policy change.

8. Do Not Track. Polisis did not report any assessments from this group. Evidently, privacy policies in our corpora never mentioned the term. This observation confirms that support for Do Not Track signals is not widely adopted by privacy policies.

9. International & Specific Audiences. None of the E.U. government privacy policies studied had a specific section on *children's privacy*. U.S. agencies, too, have a lower percentage of dedicating a section to *children's privacy*, as confirmed by PrivacyCheck.

10. Other. We were able to categorize all the reported assessments under a category above, leaving none for this category.

4.3 Availability

We make publicly available our data, including links to all the privacy policies, their downloaded text used in the experiments, the results of running the tools on them, and the SPSS outputs.

A public GitHub repository at https://github.com/nokhbehzaeem/GovVsCompanies contains our data. The structure of this repository is as follows. The Corpora folder includes five text files, each containing the URLs of privacy policies of one corpus. This folder also includes five folders, each containing the downloaded html files of privacy policies of a corpus. The Results folder has two sub-folders, named after PrivacyCheck and Polisis. Each of these folders contains a tab-delimited text file with the results of running the respective tool on all the five corpora. This text file is the input to SPSS. The output of analyzing data with SPSS, to run the chi-square test and calculate Cramer's V, is provided as a pdf file.

Polisis and Privacy check are made available by their respective creators at https://pribot.org and https://identity.utexas.edu/privacycheck-for-google-chrome.

5 THREATS TO VALIDITY

The major threat to the internal validity of this study is the limitation our work inherits from the tools it uses: the level of accuracy in automatic privacy analysis. The F-1 score of Polisis ranges between 0.71 and 0.97 across its categories, with an average of 0.84 (Harkous et al., 2018). The accuracy of PrivacyCheck as reported in its original paper (Zaeem et al., 2018) ranges between 0.40 to 0.73 across its ten questions, with an average of 0.55, but the authors have improved the average accuracy to 0.60 since (Nokhbeh Zaeem et al., 2020).

To address this threat, we sought to utilize the best *available* tools. We used two different, most-recent, privacy analysis tools from independent research groups. We found the results of these tools consistent with one another.

The main threat to the external validity of our work pertains to its applicability to other privacy policies outside the set of our five corpora. We purposefully obtained three large corpora through three different routes (stock market sampling of companies and their policies, a web search, and a corpus of mobile app privacy policies) to diversify our selection of policies. We furthermore crawled the official websites of federal U.S. agencies and E.U. government agencies to collect all the corresponding privacy policies.

6 RELATED WORK

To our knowledge, this is the first work to compare privacy policies of companies and government agencies. In this section, we briefly cover other privacy policy analysis tools, besides Polisis and PrivacyCheck, which we already discussed in details. Meanwhile, we also recap some of the work in the literature on assembling privacy policy corpora.

Privee (Zimmeck and Bellovin, 2014) is an older automatic privacy policy analysis tool. Building on the crowd sourcing privacy analysis framework ToS;DR (ToS;DR, 2012), Privee combines crowd sourcing with rule and machine learning classifiers to classify privacy policies that are not already rated in the crowd sourcing repository.

The Usable Privacy Project[6] (Sadeh et al., 2013) takes advantage of machine learning and crowd sourcing to semi-automatically annotate privacy policies. This project annotates (Wilson et al., 2016b; Wilson et al., 2016a) a corpus of 115 policies with attributes and data practices, the same corpus that Polisis uses to extract its coarse- and fine-grained classes.

We covered Polisis (Harkous et al., 2018) in Section 2.2. Pribots (Harkous et al., 2016) is from the same authors and is a chat bot that answers questions about privacy policies. Polisis and Pribots build upon a corpus of 13,000 mobile app policies. We based our corpus of 100 mobile app policies on this corpus.

MAPS (Zimmeck et al., 2019) scales mobile app privacy analysis to more than one million apps. Their corpus of 350 human-annotated policies is publicly available. Similar studies of mobile app privacy policies are on the rise (Zimmeck et al., 2016).

Other researchers, too, have applied machine learning and natural language processing in privacy policy analysis (Fawaz et al., 2019). PolicyLint (Andow et al., 2019) is a natural language processing tool that identifies potential contradictions that may arise inside the same privacy policy. PolicyLint is tested on a corpus of 11,430 privacy policies from mobile apps. This corpus, however, is not made public. PrivacyGuide (Tesfay et al., 2018) is a machine learning and natural language processing tool inspired by the GDPR. It uses a corpus of 45 policies from the most accessed websites in Europe. PrivacyGuide, however, is not publicly available, as opposed to PrivacyCheck and Polisis.

The Center for Identity at the University of Texas at Austin[7] targets many aspects of identity management and privacy (Zaeem et al., 2017; Zaeem et al., 2016a; Zaeem et al., 2016b; Zaiss et al., 2019; Rana et al., 2019). They developed PrivacyCheck (Zaeem et al., 2018), as detailed in Section 2.1. The same research group studied privacy policies across industries (Zaeem and Barber, 2017) and produced a privacy policy corpus of 400 companies. We obtained our stock market and web search corpora from them.

7 CONCLUSION

Through the application of two machine learning privacy analysis tools (Polisis and PrivacyCheck) on five corpora of privacy policies (including policies of 285 U.S. and E.U. government agencies and 550 companies) we uncovered (with a significance of 99.99%) the differences that exist between privacy policies of government agencies and companies. We measured and reported Cramer's V, a chi-square-based measure of association as well. The results of the two machine learning tools were consistent and confirmed the common expectation that government agencies are better in not collecting and protecting user data, including sensitive financial information, Social Security Numbers, and user location. We, however, uncovered some unexpected results too. For example, many of the government agencies lack a separate section on children's privacy or detailed security measures taken to protect user data. Our experiments demonstrated how privacy policies of European government agencies perform better than their U.S. counterparts, with respect to notifying users of policy change, giving users the right to edit/delete their data, and limiting data retention. Our work quantifies the actual differences between corporate and government privacy policies and compares U.S. and E.U. policies together. By making our corpora and results publicly available, we hope that this work also assists the research community in investigating privacy policies and enhancing them.

[6] https://usableprivacy.org

[7] https://identity.utexas.edu

ACKNOWLEDGMENTS

We thank Hamza Harkous and the Polisis authors for creating a Polisis API and providing their data.

REFERENCES

Andow, B., Mahmud, S. Y., Wang, W., Whitaker, J., Enck, W., Reaves, B., Singh, K., and Xie, T. (2019). Policylint: investigating internal privacy policy contradictions on Google play. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 585–602.

Ermakova, T., Baumann, A., Fabian, B., and Krasnova, H. (2014). Privacy policies and users' trust: Does readability matter? In *20th Americas Conference on Information Systems (AMCIS)*.

Fawaz, K., Linden, T., and Harkous, H. (2019). The applications of machine learning in privacy notice and choice. In *2019 11th International Conference on Communication Systems & Networks (COMSNETS)*, pages 118–124. IEEE.

FTC (2000). Privacy online: Fair information practices in the electronic marketplace: A federal trade commission report to congress.

Graber, M. A., D Alessandro, D. M., and Johnson-West, J. (2002). Reading level of privacy policies on internet health web sites. *Journal of Family Practice*, 51(7):642–642.

Harkous, H., Fawaz, K., Lebret, R., Schaub, F., Shin, K. G., and Aberer, K. (2018). Polisis: Automated analysis and presentation of privacy policies using deep learning. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 531–548.

Harkous, H., Fawaz, K., Shin, K. G., and Aberer, K. (2016). Pribots: Conversational privacy with chatbots. In *Twelfth Symposium on Usable Privacy and Security (SOUPS) 2016*.

Linden, T., Harkous, H., and Fawaz, K. (2018). The privacy policy landscape after the gdpr. *arXiv preprint arXiv:1809.08396*.

McDonald, A. M. and Cranor, L. F. (2008). the cost of reading privacy policies. *I/S: A Journal of Law and Policy for the Information Society*, 4:543.

Milne, G. R., Culnan, M. J., and Greene, H. (2006). A longitudinal assessment of online privacy notice readability. *Journal of Public Policy & Marketing*, 25(2):238–249.

Nokhbeh Zaeem, R., Anya, S., Issa, A., Nimergood, J., Rogers, I., Shah, V., Srivastava, A., and Barber, K. S. (2020). Privacycheck v2: A tool that recaps privacy policies for you. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 3441–3444.

Rana, R., Zaeem, R. N., and Barber, K. S. (2019). An assessment of blockchain identity solutions: Minimizing risk and liability of authentication. In *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 26–33.

Regard, H. (1980). Recommendation of the council concerning guidelines governing the protection of privacy and transborder flows of personal data.

Sadeh, N., Acquisti, A., Breaux, T. D., Cranor, L. F., McDonalda, A. M., Reidenbergb, J. R., Smith, N. A., Liu, F., Russellb, N. C., Schaub, F., et al. (2013). The usable privacy policy project. Technical report, Technical Report, CMU-ISR-13-119, Carnegie Mellon University.

Tesfay, W. B., Hofmann, P., Nakamura, T., Kiyomoto, S., and Serna, J. (2018). Privacyguide: towards an implementation of the EU GDPR on internet privacy policy evaluation. In *Proceedings of the Fourth ACM International Workshop on Security and Privacy Analytics*, pages 15–21. ACM.

ToS;DR (2012). Terms of service; didn't read.

Union, E. European union law.

Wilson, S., Schaub, F., Dara, A. A., Liu, F., Cherivirala, S., Leon, P. G., Andersen, M. S., Zimmeck, S., Sathyendra, K. M., Russell, N. C., et al. (2016a). The creation and analysis of a website privacy policy corpus. In *Annual Meeting of the Association for Computational Linguistics*, pages 1330–13340.

Wilson, S., Schaub, F., Ramanath, R., Sadeh, N., Liu, F., Smith, N. A., and Liu, F. (2016b). Crowdsourcing annotations for websites' privacy policies: Can it really work? In *Proceedings of the 25th International Conference on World Wide Web*, pages 133–143.

Zaeem, R. N. and Barber, K. S. (2017). A study of web privacy policies across industries. *Journal of Information Privacy and Security*, 13(4):169–185.

Zaeem, R. N. and Barber, K. S. (2020). The effect of the GDPR on privacy policies: Recent progress and future promise. *ACM Transactions on Management of Information Systems*.

Zaeem, R. N., Budalakoti, S., Barber, K. S., Rasheed, M., and Bajaj, C. (2016a). Predicting and explaining identity risk, exposure and cost using the ecosystem of identity attributes. In *2016 IEEE International Carnahan Conference on Security Technology (ICCST)*, pages 1–8. IEEE.

Zaeem, R. N., German, R. L., and Barber, K. S. (2018). Privacycheck: Automatic summarization of privacy policies using data mining. *ACM Transactions on Internet Technology (TOIT)*, 18(4):53.

Zaeem, R. N., Manoharan, M., and Barber, K. S. (2016b). Risk kit: Highlighting vulnerable identity assets for specific age groups. In *2016 European Intelligence and Security Informatics Conference (EISIC)*, pages 32–38. IEEE.

Zaeem, R. N., Manoharan, M., Yang, Y., and Barber, K. S. (2017). Modeling and analysis of identity threat behaviors through text mining of identity theft stories. *Computers & Security*, 65:50–63.

Zaiss, J., Nokhbeh Zaeem, R., and Barber, K. S. (2019). Identity threat assessment and prediction. *Journal of Consumer Affairs*, 53(1):58–70.

Zimmeck, S. and Bellovin, S. M. (2014). Privee: An architecture for automatically analyzing web privacy policies. In *23rd USENIX Security Symposium (USENIX Security 14)*, pages 1–16, San Diego, CA. USENIX Association.

Zimmeck, S., Story, P., Smullen, D., Ravichander, A., Wang, Z., Reidenberg, J., Russell, N. C., and Sadeh, N. (2019). Maps: Scaling privacy compliance analysis to a million apps. *Proceedings on Privacy Enhancing Technologies*, 2019(3):66–86.

Zimmeck, S., Wang, Z., Zou, L., Iyengar, R., Liu, B., Schaub, F., Wilson, S., Sadeh, N., Bellovin, S., and Reidenberg, J. (2016). Automated analysis of privacy requirements for mobile apps. In *2016 AAAI Fall Symposium Series*.