The Effect of the GDPR on Privacy Policies: Recent Progress and Future Promise

RAZIEH NOKHBEH ZAEEM and K. SUZANNE BARBER, The Center for Identity, The University of Texas at Austin

The General Data Protection Regulation (GDPR) is considered by some to be the most important change in data privacy regulation in 20 years. Effective May 2018, the European Union GDPR privacy law applies to any organization that collects and processes the personal information of EU citizens within or outside the EU. In this work, we seek to quantify the progress the GDPR has made in improving privacy policies around the globe. We leverage our data mining tool, PrivacyCheck, to automatically compare three corpora (totaling 550) of privacy policies, pre- and post-GDPR. In addition, to evaluate the current level of compliance with the GDPR around the globe, we manually studied the policies within two corpora (450 policies). We find that the GDPR has made progress in protecting user data, but more progress is necessary—particularly in the area of giving users the right to edit and delete their information—to entirely fulfill the GDPR's promise. We also observe that the GDPR encourages sharing user data with law enforcement, and as a result, many policies have facilitated such sharing after the GDPR. Finally, we see that when there is non-compliance with the GDPR, it is often in the form of failing to explicitly indicate compliance, which in turn speaks to an organization's lack of transparency and disclosure regarding their processing and protection of personal information. If Personally Identifiable Information (PII) is the "currency of the Internet," these findings mark continued alarm regarding an individual's agency to protect and secure their PII assets.

CCS Concepts: • **Security and privacy → Privacy protections**;

Additional Key Words and Phrases: Privacy, policy, GDPR, PrivacyCheck

ACM Reference format:
Razieh Nokhbeh Zaeem and K. Suzanne Barber. 2020. The Effect of the GDPR on Privacy Policies: Recent Progress and Future Promise. *ACM Trans. Manage. Inf. Syst.* 12, 1, Article 2 (December 2020), 20 pages.
https://doi.org/10.1145/3389685

1 INTRODUCTION

The General Data Protection Regulation (GDPR) represents the most significant data protection and privacy regulation in many years. While the GDPR is a European Union law, it covers any organization that collects or processes EU citizen data independent of the organization's location. Due to the global nature of commerce and people's movements, the GDPR drove businesses around

This work was in part funded by the Center for Identity's Strategic Partners. The complete list of Partners can be found at https://identity.utexas.edu/strategic-partners.
Authors' addresses: R. N. Zaeem and K. S. Barber, The Center for Identity, The University of Texas at Austin, 201 E 24th St, Austin, TX 78712; emails: {razieh, sbarber}@identity.utexas.edu.

2158-656X/2020/12-ART2 $15.00
https://doi.org/10.1145/3389685

the world to make important decisions and changes regarding how they collect and process their employees' and customers' Personally Identifiable Information (PII).

The main goals of the GDPR are to give individuals control over their personal data and to unify the regulation within the EU to facilitate business. Its key principles are:

- Lawfulness, fairness, and transparency
- Purpose limitation
- Data minimization
- Accuracy
- Storage limitation
- Integrity and confidentiality (security)
- Accountability

The GDPR went into effect on May 25, 2018. As a result, many companies that do business completely or partially in the EU or handle EU citizens' data updated their online privacy policies around the same time in order to comply with the GDPR [4]. It is also important to note that the GDPR has already motivated other advancements in privacy regulation and continues to do so around the globe as consumers demand their data rights.[1] Evaluating the actual effect of the GDPR on online privacy policies of companies is a significant research question, which has not, as we review in Section 2, received the attention it deserves. It is critical to quantify both the progress the GDPR has made toward improving privacy policies and the remaining work to be done in those policies to fulfill the promise of the GDPR in protecting consumer rights.

In this work, we assess the *progress* made in privacy policies by comparing pre- and post-GDPR policies. We seek to measure the change in privacy policies in an automated and scalable manner through the use of data mining. Therefore, we utilize our proprietary data mining tool PrivacyCheck[2] [13, 24]—developed at the Center for Identity at the University of Texas at Austin (UT CID)—that answers 10 fundamental questions concerning the privacy and security of user data in privacy policies. These 10 questions were compiled from previous work on privacy factors, for example, the work of the Organization for Economic Co-operation and Development [15], the work of the Federal Trade Commission (FTC) including those on Fair Information Practice Principles (FIPPs) [6–8], and other academic [22] and business research work [16].

Furthermore, we quantify the improvements that still need to be made in privacy policies, particularly with US-based companies, to completely meet the requirements of the GDPR and accomplish its *promise*. As others have found [4], companies and organizations that do business in the EU have moved toward complying with the GDPR to avoid its fines and sanctions. Many privacy policies, however, still violate several key GDPR requirements [11]. The non-compliance might be even greater in policies of companies based outside the EU, for example, in the US. In addition to overarching privacy concerns measured by PrivacyCheck, we construct another list of 10 questions directly based on the GDPR requirements and manually review privacy policies to judge their degree of compliance.

We take advantage of three corpora of online privacy policies of 400, 50, and 100 companies. We compiled the first corpus of 400 policies by considering 10% of all the companies listed on the NYSE, Nasdaq, and AMEX stock markets [13, 24]. We selected the second corpus of 50 privacy policies from popular companies through a web search [13, 24]. Finally, we obtained the third corpus of 100 policies from the authors of Polisis [9], who used a crawler to search the web for privacy policies. Using the total of 550 policies, we make two contributions:

[1] California Consumer Privacy Act (CCPA): https://www.caprivacy.org.
[2] Available online at https://identity.utexas.edu/privacycheck-for-google-chrome.

(1) We apply our own data mining tool, PrivacyCheck, to compare the privacy policies before and after the GDPR through PrivacyCheck's 10 questions to assess the impact of the GDPR.
(2) We manually study these privacy policies after the GDPR went into effect to measure how close the policies are to full compliance with the GDPR. We distill another 10 questions, previously not supported by PrivacyCheck, directly from the GDPR.

2 RELATED WORK

Very few researchers have studied the actual effect of the GDPR on privacy policies. While some researchers have looked at privacy policies to build corpora for machine learning, they either do not report anything about the corpus itself [18, 19] or have very small corpora (e.g., from 14 [3] to 50 policies [12]).

Degeling and colleagues are among the few who studied a sizable set of privacy policies (from more than 6,500 websites) across the EU [4] with the help of automation and compared policies before and after the GDPR. That work, however, focused exclusively on the consent to use cookies in privacy policies. Libert et al. [10] studied privacy policies of seven countries of the EU before and after the GDPR went into effect but also focused on cookies only.

To the best of our knowledge, the most closely related work comes from Linden et al. [11], which examines policies from both inside and outside the EU. While they use over 6,000 privacy policies, with the exception of the visual representation evaluation that is done by Amazon MTurk users, the rest of the compliance assessment is done automatically. The automation is in turn based on Polisis [9], a deep learning tool developed by the authors. They concluded that, even though the GDPR has prompted a general overhaul in privacy policies, many policies still do not meet several GDPR requirements. The major difference of our work from that of Linden et al. [11] is that our evaluation is a more granular measurement—particularly with 10 new questions aimed at the heart of the GDRP. Another advantage of our work is that we assess privacy policies both manually and automatically. We manually and thoroughly investigate 450 unique privacy policies and ensure integrity by manually checking a third of these policies twice, reporting on the consistency of our manual evaluation. We further automatically explore privacy policies through data mining. Contissa et al. [3] compare pre- and post-GDPR privacy policies with a data mining tool, but their work is limited to only 14 policies.

3 BACKGROUND: PRIVACY POLICIES AND THE GDPR

Privacy policies have become the de facto way of communicating how a company or organization—and particularly its website—collect, share, and use PII. Many government agencies around the globe (e.g, the FTC in the US) mandate posting privacy policies. Furthermore, many seek to protect consumers' PII by enforcing laws and regulations on these policies.

One of the latest of such regulations and laws is the GDPR [21] in the EU. As the GDPR defines, personal data means "any information relating to an identified or identifiable natural person" such as identification numbers, location data, or physical data, among others. Considering the protection of personal data a fundamental right of natural persons, the GDPR (in its Article 5) mandates that personal data shall be

(1) processed lawfully, fairly, and with transparency.
(2) collected for explicitly specified and limited purposes.
(3) adequate, relevant, and only minimized to what is necessary.
(4) accurate and up to date.

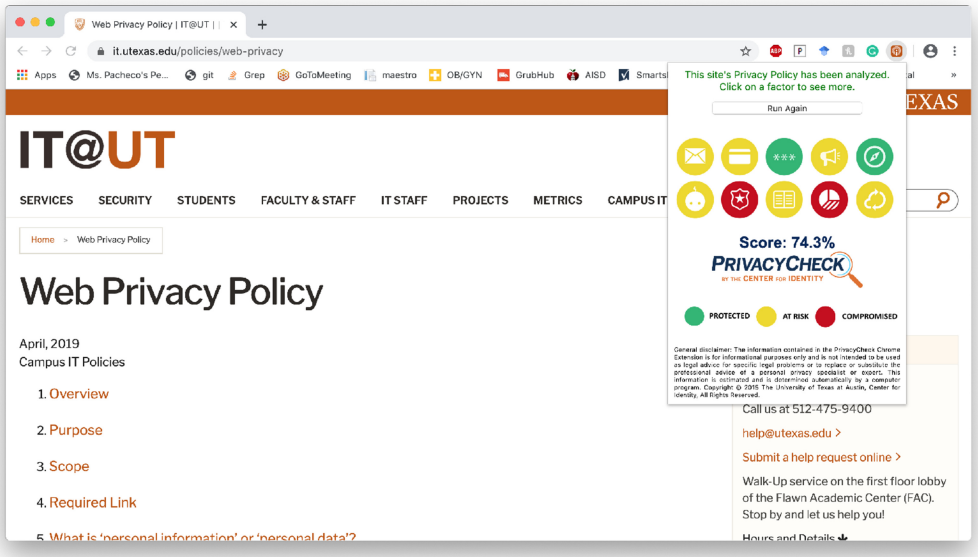

Fig. 1. A screenshot of running PrivacyCheck on the web privacy policy of the University of Texas at Austin.

(5) kept such as permits identification for no longer than necessary (also known as storage limitation).
(6) processed with integrity and confidentiality.

Furthermore, the controller of data is held accountable for complying with the above requirements.

The EU passed the GDPR in April 2016 and started its 2-year grace period. After 2 years, the GDPR went into full effect on May 25, 2018. Companies and organizations that handle information of EU citizens were required to comply with the GDPR by this date. As other researchers have noted [4], these companies sought to change their privacy policies in accordance with the GDPR to avoid its fines and sanctions. Nonetheless, many are still not in full compliance with the GDPR [11].

4 BACKGROUND: PRIVACYCHECK

PrivacyCheck [13, 24] is our data mining tool that analyzes the text of privacy policies and answers 10 basic questions concerning the privacy and security of user data, what information is gathered from them, and how this information is used. PrivacyCheck is available as browser plug-ins for Chrome, Firefox, and Safari.[3] Figure 1 shows a screenshot of running PrivacyCheck on the web privacy policy of the University of Texas at Austin.

PrivacyCheck analyzes the privacy policy text to find out if and how it

(1) handles the user's email address.
(2) handles the user's credit card number and home address.
(3) handles the user's Social Security number.
(4) uses or shares PII for marketing purposes.
(5) tracks or shares location.

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

Table 1. Risk Levels that PrivacyCheck Assigns to Privacy Policies for Its 10 Privacy Factors

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

(6) collects PII from children under 13.
(7) shares personal information with law enforcement.
(8) notifies the user when the privacy policy changes.
(9) allows the user to control their data by editing or deleting information.
(10) collects or shares aggregated data related to personal information.

PrivacyCheck scores the privacy policy for each of these factors according to three levels of risk: red (high), yellow (medium), and green (low). The user of PrivacyCheck can read more about each factor and the assigned risk level by hovering over each of the 10 icons (Figure 1).

Table 1 (from our previous work on PrivacyCheck [24]) shows the risk levels for each of the privacy factors. These 10 factors were compiled from previous work, for example, the work of the Organization for Economic Co-operation and Development [15], the work of the FTC [6–8], and other academic [22] and business research work [16].

4.1 PrivacyCheck's Architecture

PrivacyCheck utilizes a client-server architecture. The browser extension client pre-processes the privacy policy's text and sends the processed text to the data mining server. We have trained 10 Naive Bayes classifier models running at the server to answer each of the 10 questions of Table 1. Finally, the server sends these answers to the PrivacyCheck client to display as colorful icons accompanied by short text snippets.

4.2 PrivacyCheck's Client-Side Text Processing

In order to extract parts of privacy policies that are related to each of the privacy factors, PrivacyCheck runs Algorithm 1. This algorithm breaks the text into paragraphs, removes punctuation, converts uppercase to lowercase, removes stop words, performs stemming, and keeps only the paragraphs that have at least one keyword related to a particular factor. The input to this algorithm is the text from the web page T, a set of privacy factors F that we would like to consider, and the set of keywords K_f for each factor f. The algorithm's output is a text snippet S_f for each of these factors. Line 1 breaks the web page text into paragraphs. Then, for each paragraph, the algorithm performs the following. Line 3 replaces all non-alphanumeric characters with spaces, effectively removing all punctuation marks to create the **p**unctuation-**l**ess paragraph *pl*. Line 4 converts this punctuation-less paragraph to **l**ower**c**ase *lc*. The next line removes all the stop words (**s**top word-**l**ess *sl*) and replaces any sequence of spaces (generated through previous text manipulating lines or originally present in the text) with one single space. Finally, Line 6 performs stemming to keep

only the word roots in the **f**inal **p**aragraph fp. Line 7 puts those word roots in W. Line 8 iterates through the factors and for each factor does the following steps. If any word of this paragraph contains any keyword for the factor, then the entire paragraph fp is kept in S_f for that factor and the algorithm moves on to the next factor. Finally, after iterating over each factor and over each paragraph, the algorithm returns S, which contains all S_fs.

ALGORITHM 1: Text pre-processing

input: Web Page Text T, Set of Privacy Factors F, Set of Keywords for Each Factor K_f
output: Text Snippet for Each Privacy Factor S_f

```
1  P ← T.split('\n')
2  foreach paragraph p in P do
3      pl ← p.replace(/[^A-Za-z0-9-]/g,'''') // punctuation-less removes any non alphanumeric
           character
4      lc ← pl.toLowerCase() // converts to lowercase
5      sl ← lc.replace(/\b(i|me|my|...|should|now)\b/g,' ').replace(/\ s{2,}/g,'''') // stopword-less
           removes stop words and replaces any double or more spaces with a single space
6      fp ← sl.stem() // final paragraph includes only the word stems
7      W ← fp.split(' ')
8      foreach factor f in F do
9          nextFactor:
10         foreach word w in W do
11             foreach keyword k in K_f do
12                 if w.contains(k) then
13                     S_f ← S_f + fp+' '
14                     break nextFactor
15                 end
16             end
17         end
18     end
19 end
20 return S
```

4.3 PrivacyCheck's Server-Side Data Mining Models

The current version of PrivacyCheck utilizes data mining models—particularly Naive Bayes classifiers from Scikit-Learn—to summarize any HTML page that contains a privacy policy. We trained the Naive Bayes classifiers of PrivacyCheck on a corpus of 400 privacy policies and their manually identified risk levels. This corpus is the same as one of our corpora used in this work to evaluate the GDPR landscape and is explained in detail in Section 5.1.

To train the models, we leveraged the 400 privacy policies. A team of graduate and undergraduate students read all of these policies and scored each policy according to Table 1, using the red/yellow/green risk levels. We performed quality control by assigning 15% of the privacy policies (60 policies, randomly selected) to two team members in order to train them to be consistent in assigning risk levels.

Next, we applied the pre-processing algorithm to process the 400 policies and put together a training file for each factor. The training file includes the corresponding text snippet and the manually determined risk levels for each of the 400 policies. We trained a Naive Bayes classification model against each file, the independent variable being the text snippet and the dependent variable

Table 2. Classification Accuracy of Models (F_1 Score)

Factor	F_1 Score
(1) Email Address	0.76
(2) Credit Card Number	0.59
(3) Social Security Number	0.68
(4) Ads and Marketing	0.57
(5) Location	0.62
(6) Collecting PII of Children	0.59
(7) Sharing w/ Law Enforcement	0.57
(8) Policy Change	0.59
(9) Control of Data	0.48
(10) Data Aggregation	0.68

Table 3. PrivacyCheck Fivefold Cross-Validation Precision Results: The Percentage of PrivacyCheck Correctly Predicting Ground Truth

	Email Address	Credit Card Number	Social Security Number	Ads and Marketing	Location	Collecting PII of Children	Sharing w/ Law Enforcement	Policy Change	Control of Data	Data Aggregation
Iteration 1	60%	43%	61%	48%	65%	36%	43%	49%	33%	88%
Iteration 2	63%	45%	50%	55%	43%	49%	45%	41%	40%	56%
Iteration 3	90%	54%	58%	64%	60%	56%	53%	66%	44%	69%
Iteration 4	66%	44%	48%	45%	48%	59%	50%	61%	46%	60%
Iteration 5	85%	70%	65%	54%	74%	54%	45%	56%	36%	64%
Average	73%	51%	56%	53%	58%	51%	47%	55%	40%	67%

being the class (i.e., the risk level). Each privacy policy had one feature for every model: the text snipped generated by the text pre-processing algorithm for the corresponding privacy factor. We achieved classification accuracies (F_1 scores) ranging from 0.48 to 0.76 for various privacy factors (as shown in Table 2).

In order to estimate the prediction ability of the models, we performed fivefold cross-validation using the corpus of 400 privacy policies. In each iteration of cross-validation, the 10 models were trained using 320 privacy policies and then tested against the remaining 80 policies. The ground truth, as always, was the result of manually investigating the policy. The models matched the ground truth 40% to 73% of the time, as shown in Table 3.

4.4 Tools Similar to PrivacyCheck

In this section, we briefly cover other privacy policy analysis tools similar to PrivacyCheck. We have developed PrivacyCheck [24] and used it to study privacy policies across industries [13] at

the Center for Identity at the University of Texas at Austin,[4] where we target many aspects of identity management and privacy [14, 23, 25–27].

The Usable Privacy Project[5] [17] takes advantage of natural language processing, machine learning, privacy preference modeling, crowd-sourcing, and formal methods to semi-automatically annotate privacy policies. Most recently, the Usable Privacy Project released Polisis [9]. Polisis includes 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. At its core, Polisis is a neural network classifier trained on privacy policies retrieved from the Google Play store. Polisis was developed in 2018, 3 years after PrivacyCheck.

Privee [28] is an older automatic privacy policy analysis tool by some of the authors of Polisis. Building on the crowd-sourcing privacy analysis framework ToS;DR [20], Privee combines crowd-sourcing with rule and machine learning classifiers to classify privacy policies that are not already rated in the crowd-sourcing repository.

Other researchers too have applied machine learning and natural language processing in privacy policy analysis [5]. PolicyLint [1] is a natural language processing tool that identifies potential contradictions that may arise inside the same privacy policy. PrivacyGuide [19] 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.

In this article, we utilize PrivacyCheck to quantify how privacy policies changed after (versus before) the GDPR. We also report on the current state of the art of these policies with respect to the 10 factors of PrivacyCheck. We use multiple corpora to evaluate the current privacy landscape with respect to the GDPR. The next section elaborates on these corpora.

5 CORPORA

We use three corpora of online privacy policies of 400, 50, and 100 companies, respectively. We made sure that there is no overlap between the three corpora.

5.1 Corpus of 400 Privacy Policies

We compiled the first corpus of 400 policies by considering 10% of all the companies listed on NYSE, Nasdaq, and AMEX stock markets [13, 24]. In April 2016, while training PrivacyCheck, we assembled a corpus of 400 privacy policies across industries by randomly selecting 10% of all the companies listed by these stock markets. Since this corpus was the training set of PrivacyCheck, the risk levels for each of the PrivacyCheck factors were manually assigned. This manual assignment guarantees that the risk levels for this corpus are accurate for the policies before the GDPR. Recall that the GDPR was passed in April 2016, the same month in which we trained PrivacyCheck.

5.2 Corpus of 50 Privacy Policies

We selected the second corpus of 50 privacy policies from popular companies through a web search [13, 24]. As a part of testing PrivacyCheck, we performed a Google search with terms "privacy policy" and selected the first 50 non-sponsored search results that we had not used in the training phase. The set of privacy policies included well-known websites (e.g., Google, Facebook, Twitter, CNN, Wikipedia) and less-known websites (e.g., Ello, OwnPhones, and Automattic).

For PrivacyCheck risk levels before the GDPR, we fetched the April (May when April was not available) 2016 version of the company's privacy policy from an Internet archive named the Way Back Machine [2] and ran PrivacyCheck on it.

[4] https://identity.utexas.edu.
[5] https://usableprivacy.org.

5.3 Corpus of 100 Privacy Policies

Finally, we obtained the third corpus from the authors of Polisis [9]. The Polisis project crawled the Google Play Store for privacy policies and gathered over 130,000 policies. Because the raw corpus of 130k policies we received did not include URLs to fetch pre- and post-GDPR versions, we automatically searched the web for the corresponding companies. To our surprise, we found that very few unique texts of policies (161) existed in the corpus, while many policy texts were redundant and repetitive. We did reach out to the authors of Polisis and they confirmed that duplication stems from apps using templates or referring to similar policy documents. Furthermore, not every company/app for the policies of this corpus existed in May 2016 and had a privacy policy before the GDPR. We were able to find 100 *unique* privacy policies and their companies' links to before/after the GDPR version.

For PrivacyCheck risk levels before the GDPR, we fetched the April/May 2016 version of the policy from the Way Back Machine and ran PrivacyCheck.

For all three corpora, to record the PrivacyCheck risk levels after the GDPR, we ran PrivacyCheck on each privacy policy in August 2019 and saved the results.

Using the three corpora, we report multiple sets of results: (1) how privacy policies have changed after the GDPR versus what they were before it, according to the risk levels measured by PrivacyCheck; (2) the state of the art of privacy policies after the GDPR, according to PrivacyCheck; and (3) the level of compliance of privacy policies with respect to the GDPR, evaluated through manual investigation of policies. Because the PrivacyCheck factors *before the GDPR* were measured with different accuracy for different corpora (recall that the corpus of 400 policies was PrivacyCheck's training set), we outline the pre- and post-GDPR comparison by corpus. For the rest of the results (the state of the art of policies according to PrivacyCheck and the manual examination of compliance with the GDPR), we combine all the three corpora into one corpus of 550 policies.

6 PRIVACY POLICIES BEFORE VERSUS AFTER THE GDPR ACCORDING TO PRIVACYCHECK

First, we compare the risk levels that PrivacyCheck reports for every policy before (2016) and after (2019) the GDPR.

6.1 Corpus of 400 Privacy Policies

Out of the 400 policies of this corpus, 138 policies did not change at all with respect to PrivacyCheck. In 99 policies, only one of the PrivacyCheck factors changed, out of which, in 62 policies, the risk level improved (e.g, from the yellow level to the green level, or from the red level to the yellow level) after the GDPR.

Considering the PrivacyCheck factors individually, this corpus shows the positive effect of the GDPR on protecting the privacy of children. With 281 out of 400 (70%) policies at the same level of protection for children before and after the GDPR, a significant fraction of 29% (or 114 policies) in this corpus improved their protection of children's privacy. For the data aggregation factor, 76% (303 policies) get the same score from PrivacyCheck before and after the GDPR, but 17% (69 policies) get a better score. Social security and credit card numbers most commonly receive the same level of privacy protection before and after the GDPR. Table 4 shows the number of policies in this corpus that diminished, did not change, or improved with respect to each of the factors.

6.2 Corpus of 50 Privacy Policies

In this corpus, 19 out of 50 policies did not change at all with respect to what PrivacyCheck reports. Overall, in this corpus, the policies did not change significantly after the GDPR when considering

Table 4. Change Measured by PrivacyCheck Factors for the Corpus of 400 Policies

#Policies	Email Address	Credit Card Number	Social Security Number	Ads and Marketing	Location	Collecting PII of Children	Sharing w/ Law Enforcement	Policy Change	Control of Data	Data Aggregation
Diminished	12	16	1	35	4	5	31	22	55	28
Did Not Change	357	346	370	332	345	281	327	326	335	303
Improved	31	38	29	33	51	114	42	52	10	69

Table 5. Change Measured by PrivacyCheck Factors for the Corpus of 50 Policies

#Policies	Email Address	Credit Card Number	Social Security Number	Ads and Marketing	Location	Collecting PII of Children	Sharing w/ Law Enforcement	Policy Change	Control of Data	Data Aggregation
Diminished	0	8	0	1	5	0	5	1	10	0
Did Not Change	48	37	49	46	40	48	35	45	38	49
Improved	2	5	1	3	5	2	10	4	2	1

the PrivacyCheck factors. According to Table 5, many of the policies stayed at the same level of protection for most of the PrivacyCheck factors. The way policies treat "Email Address," "Social Security Number," "Ads and Marketing," "Collecting PII of Children," "Policy Change," and "Data Aggregation" remains over 90% (i.e., in at least 45 policies) the same. The GDPR does explicitly increase the age of protection for children from 13 to 16 and requires policies to notify users of the use of cookies for ads and marketing. Nonetheless, this small corpus does not reflect these changes.

6.3 Corpus of 100 Privacy Policies

More than 60% (61 policies) in this corpus did not change at all with respect to PrivacyCheck factors. The trend in this corpus is similar to that of the 50-policy corpus. The attitude toward "Email Address," "Ads and Marketing," and "Data Aggregation" mostly remains the same in the policies of this corpus, while it changes the most when it comes to "Credit Card Number" and "Control of Data," and not always for the better. We witness mixed changes in the scores that the

Table 6. Change Measured by PrivacyCheck Factors for the Corpus of 100 Policies

#Policies	Email Address	Credit Card Number	Social Security Number	Ads and Marketing	Location	Collecting PII of Children	Sharing w/ Law Enforcement	Policy Change	Control of Data	Data Aggregation
Diminished	0	11	1	3	8	3	6	4	9	1
Did Not Change	97	81	92	97	87	91	88	94	85	99
Improved	3	8	7	0	5	6	6	2	6	0

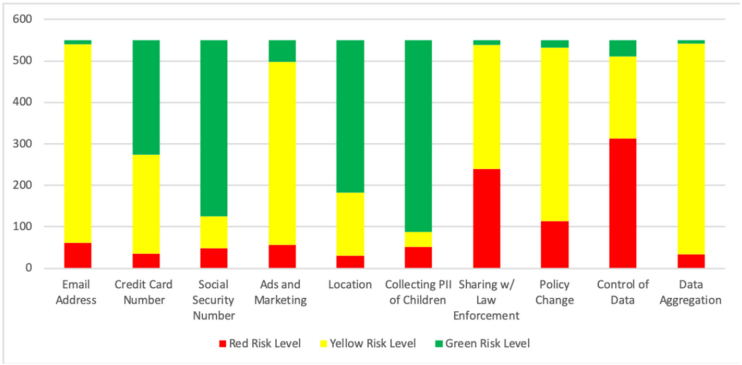

Fig. 2. State of the art of all 550 privacy policies according to PrivacyCheck factors.

policies of this corpus receive from PrivacyCheck for "Credit Card Number" and "Control of Data" (Table 6).

7 STATE OF THE ART OF PRIVACY POLICIES ACCORDING TO PRIVACYCHECK

In the second set of experiments, we analyze the risk levels that PrivacyCheck reports for every policy after the GDPR (2019) to evaluate the current state of privacy policies.

We ran PrivacyCheck on all the privacy policies of the three corpora in August 2019. Since we used the same consistent algorithm of PrivacyCheck (none of these after-GDPR policies were in the training set of PrivacyCheck), we can combine all the corpora into one corpus of 550 policies.

Figure 2 displays the number of policies that received each of the risk levels (red/yellow/green) from PrivacyCheck for each of the factors in the combined corpus of 550 policies. As the GDPR emphasizes, the children's privacy factor enjoys the highest protection in the green level. "Social Security Number" is also commonly not asked for (green level), but as Social Security Number is a US-oriented identifier, the GDPR does not explicitly address it. "Email Address" is collected but used only for the intended purpose according to many policies (the yellow level).

To elaborate more on the details, Figures 3, 4, and 5 break this analysis into the three corpora of 400, 50, and 100 policies, respectively. The general trend is the same as Figure 2.

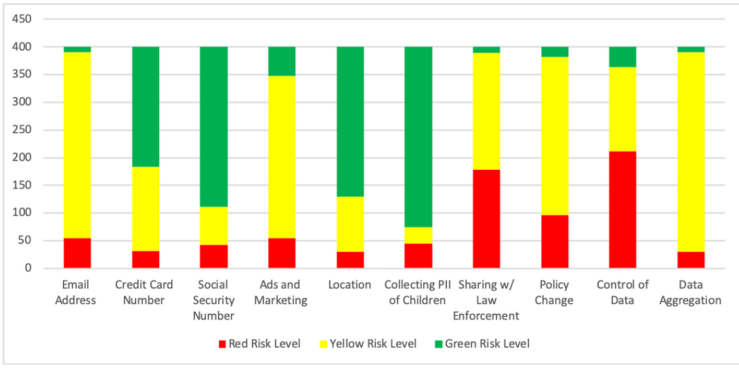

Fig. 3. State of the art of privacy policies according to PrivacyCheck factors: the corpus of 400 policies.

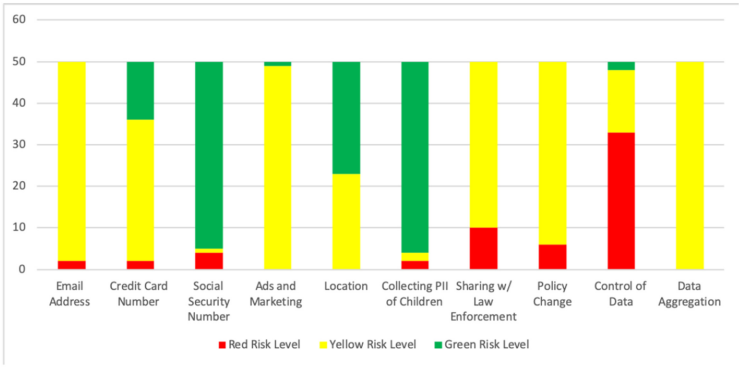

Fig. 4. State of the art of privacy policies according to PrivacyCheck factors: the corpus of 50 policies.

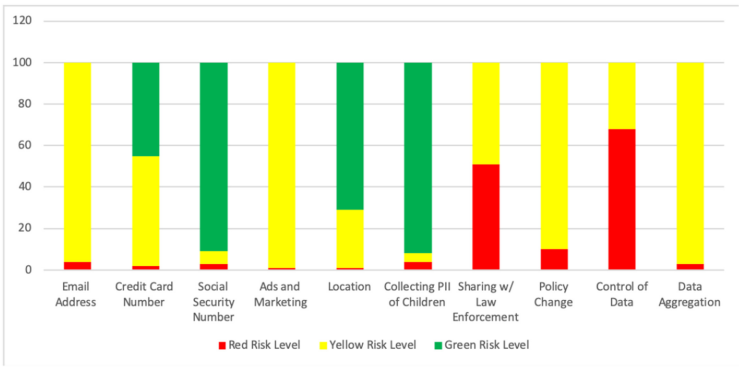

Fig. 5. State of the art of privacy policies according to PrivacyCheck factors: the corpus of 100 policies.

8 DISCUSSION OF PRIVACYCHECK FACTORS IN THE GDPR

As we compared pre- and post-GDPR policies through the lens of PrivacyCheck factors, it is appropriate to cover any discussion of these or similar privacy factors in the GDPR text[6] [21].

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

- Apart from the general protection of personal data, the GDPR does not separately address "Email Address," "Credit Card Number," and "Social Security Number."
- On the use of personal data for "Ads and Marketing," the GDPR states "Where personal data are processed for direct marketing purposes, the data subject shall have the right to object at any time to processing of personal data concerning him or her for such marketing [...]. Where the data subject objects to processing for direct marketing purposes, the personal data shall no longer be processed for such purposes." Therefore, the GDPR practically mandates the green or yellow levels for this factor, as long as the data subject (i.e., the user) is given and made aware of the option to opt out. We found that 90% of the policies we studied are at the green or yellow level for this factor.
- Besides listing one's "Location" as personal data, the GDPR explicitly mentions it when discussing and prohibiting *profiling*: "'profiling' means any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to [analyze] or predict aspects concerning that natural person's performance at work, economic situation, [...] location or movements." It appears that the GDPR accepts both green and yellow levels for the location factor and we observe that 94% of the policies studied are at either of these two levels.
- The GDPR specifically protects children "with regard to their personal data [...]. Such specific protection should, in particular, apply to the use of personal data of children for the purposes of marketing or creating personality or user profiles and the collection of personal data with regard to children when using services offered directly to a child." The GDPR modifies the definition of a child to one who is younger than 16 years old, as opposed to previous legislation that defined it as younger than 13. The vast majority of websites whose policies we analyzed do not collect children's personal data without parental consent. We used younger than 13 years of age as the definition of a child. The Children's Online Privacy Protection Act of the United States (COPPA) mandates 13, but the GDPR generally states 16.
- When it comes to "Sharing with Law Enforcement," the GDPR has two sets of guidelines. First, in its Article 23, the GDPR restricts the way the data controller is subject to the GDPR "when such a restriction respects the essence of the fundamental rights and freedoms and is [...] necessary [...] to safeguard" many aspects of law enforcement, including but not limited to public security, criminal law, and civil law. Second, in case of a data breach, in addition to notifying the data subject, "the controller should notify the personal data breach to the supervisory authority without undue delay." Based on the first set of guidelines, in essence, the GDPR deems both red and yellow risk levels for this privacy factor appropriate. It is no surprise that 98% of the policies we considered are at the red or yellow level for this factor. We discuss the second set of guidelines in the manual examination of policies.
- With regard to "Policy Change," "the request for consent shall be presented [...] in an intelligible and easily accessible form" and "[t]he data subject shall have the right to withdraw his or her consent at any time." However, the GDPR does not require the data controller to actively notify the data subject of every change in its policy. This level of protection is equal to PrivacyCheck's green level. However, very few policies in our corpora adhered to the green level. Upon manual investigation, we determined that the unclear language is the source of the apparent non-compliance and in fact many more of the policies are at the green level for this factor.
- Articles 16 and 17 of the GDPR are dedicated to the right to edit or delete personal data, the latter commonly known as the "right to be forgotten." The GDPR gives the data subject both rights but also takes into account the technical difficulties of deleting data completely as well

as the need to preserve some records to exercise legal claims. Therefore, it guarantees a more limited "right of erasure" instead of the "right to be forgotten." As a result, the GDPR accepts both green and yellow levels of PrivacyCheck for the "Control of Data" factor. Nonetheless, 57% of the policies we studied are still at the red level (not allowing the editing of data) and do not comply with the GDPR. Note that this does not necessarily mean that all those policies are currently subject to the GDPR penalties, as one of our corpora (of 400 policies) is mainly US based.

- Finally, "[t]he principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable," making both green and yellow levels justifiable for "Data Aggregation." Over 92% of the policies were at the yellow level.

9 CURRENT COMPLIANCE LEVEL OF PRIVACY POLICIES WITH THE GDPR

Researchers have found that, even after the GDPR went into full effect, many but not all privacy policies in the EU are in full compliance with it [4, 11]. In our last set of experiments, we enlist 10 important questions that indicate the level of compliance with the GDPR and manually score 450 privacy policies to see exactly how they conform with the GDPR.

Since our data mining tool, PrivacyCheck, was not trained to answer these new questions, we asked a team of students to read every policy and manually answer the questions[7] in spring 2019. The team was a class of undergraduate students in Electrical and Computer Engineering at the University of Texas at Austin. The students were asked to provide a "Yes/No/Not Discussed" answer in response to each question and provide text from the privacy policy in support of their answers. At the time of this experiment, we did not have access to the Polisis corpus. As a result, this experiment was performed on the first two corpora of 400 and 50 policies. The new questions based on the GDPR are as follows:

(1) Does the site, upon request, transfer the information a user provided to another organization of his or her choosing, where that transfer is technically feasible?
(2) Does the site advise where the company is based and where the information provided will be transferred to and processed?
(3) Does the site, upon the request of a user, deactivate the account and remove it from the active database?
(4) Does the site, in the event of a request to deactivate or remove, advise that the company will maintain a record of information in their database if it is deemed necessary for the organization to resolve any disputes, correct problems, comply with legal or regulations requirements, or enforce the Terms of Service Agreement?
(5) Does the site allow the user to object to the use of their PII or limit the way that the information is utilized? The user should again recognize that while they have a right to object to the use and/or the degree of how their information is used, their request may severely limit the service that is provided by the site.
(6) Does the site restrict the use of the PII of children under the age of 16?
(7) Does the site indicate that their data is encrypted even while at rest?
(8) Does the site ask for informed consent to data processing for one or more purposes?
(9) Does the site implement measures that meet the principles of data protection by design and by default?

[7]These answers can be later used to train PrivacyCheck for this new set of GDPR-related questions.

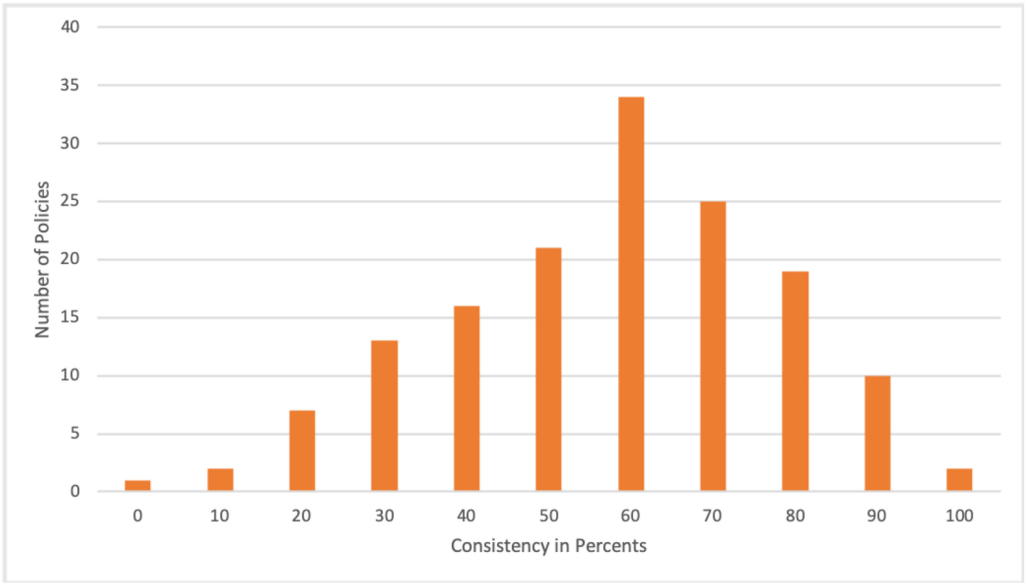

Fig. 6. Consistency between two evaluations of the same policy by different students.

(10) Does the site notify the supervisory authority without undue delay if a breach of data happens?

To assure the consistency of our manual investigation and also measure and quantify human error, we assigned 150 policies to more than one student. Figure 6 demonstrates what number of policies out of these 150 double-checked policies received exactly the same answer from two manual examiner students. For example, 34 policies had the consistency of 60% when independently graded by two students; it means that 34 policies received the exact same answer from their two examiners for 6 out of 10 GDPR-related questions, but different answers for the other 4 questions. As depicted in Figure 6, 90 out of 150 policies that were checked twice were evaluated with a consistency of at least 60%.

Figures 7 to 16 show the answers to the above GDPR-related questions among 450 policies of our first and second corpus. With only two exceptions (questions 7 and 10), the biggest group of answers to each question is "Yes" (i.e., compliance with the GDPR) and the second biggest group of answers is "Not Mentioned" (i.e., the policy does not discuss that question). For questions 7 (encrypting data while at rest) and 10 (notifying authorities in case of a breach), the majority of the policies studied do not answer the question; for the remaining policies the most answers are "Yes."

For most of the questions, only a few websites explicitly mention a policy that contradicts the GDPR. Questions 6 (protecting children under 16) and 8 (asking for informed consent) have the highest number of "No" answers. For the former, many US-based policies protect children under 13 as per US regulation, and hence do not protect children between 13 and 16 years of age. The latter shows how not all privacy policies require informed consent.

10 THREATS TO VALIDITY

There are two major threats to the internal validity of this study:

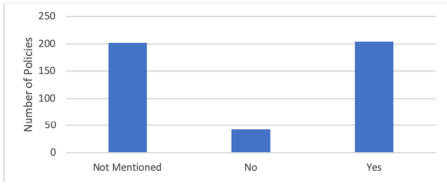

Fig. 7. GDPR question 1: Does the site, upon request, transfer the information the user provided to another organization of his or her choosing, where that transfer is technically feasible?

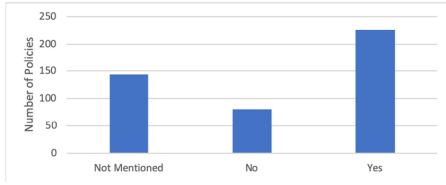

Fig. 8. GDPR question 2: Does the site advise where the company is based and where the information provided will be transferred to and processed?

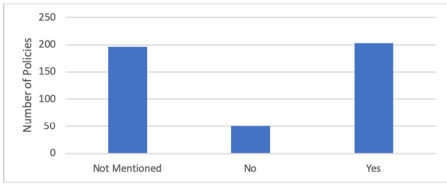

Fig. 9. GDPR question 3: Does the site, upon the request of a user, deactivate the account and remove it from the active database?

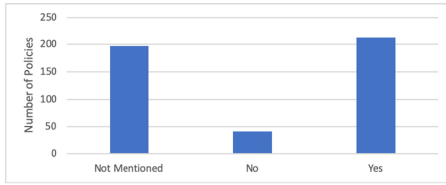

Fig. 10. GDPR question 4: Does the site, in the event of a request to deactivate or remove, advise that the company will maintain a record of information in their database if deemed necessary?

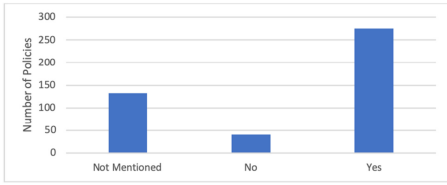

Fig. 11. GDPR question 5: Does the site allow the user object to the use of their PII or limit the way that the information is utilized?

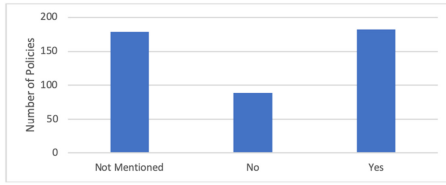

Fig. 12. GDPR question 6: Does the site restrict the use of the PII of children under the age of 16?

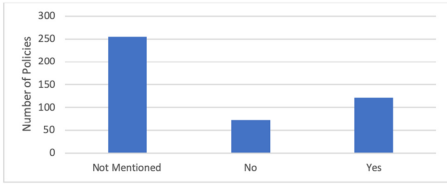

Fig. 13. GDPR question 7: Does the site advise that their data is encrypted even while at rest?

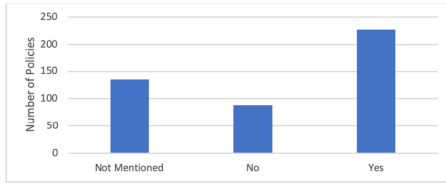

Fig. 14. GDPR question 8: Does the site ask for informed consent to data processing for one or more purposes?

(1) The potential bias in PrivacyCheck results and its data mining algorithm: To address this threat, we used the training set for some of the privacy policies that is independent of any data mining biases in PrivacyCheck. Furthermore, since the publication of the first PrivacyCheck article [24], we have improved its data mining algorithm and accuracy.

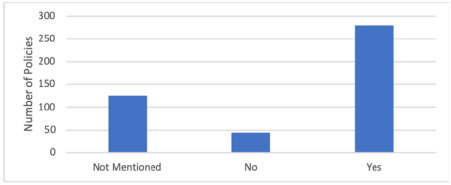

Fig. 15. GDPR question 9: Does the site implement measures that meet the principles of data protection by design and by default?

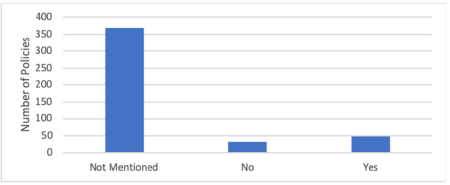

Fig. 16. GDPR question 10: Does the site notify the supervisory authority without undue delay if a breach of data happens?

(2) The potential error in manual investigation results: To address this threat, we had 150 policies independently investigated by two students and showed the consistency between these two sets of investigations.

The main threat to the external validity of our work pertains to its applicability to other privacy policies outside the set of studied policies. We purposefully obtained three large corpora through three different routes (stock market sampling of companies and their policies, a web search, and a corpus built by another research group) to diversify our selection of policies.

11 CONCLUSION AND FUTURE WORK

The EU GDPR is one of the most recent and powerful regulations passed to protect consumers' data. Not only does the GDPR give EU citizens more agency to control their own personal information with organizations inside and outside the EU, but also it has inspired sweeping new legislation in the US[8] and continues to be the most widely referenced privacy regulation as new regulations are considered in the US and around the globe.

In this work, we examined the landscape of online privacy policies to evaluate the effect of the GDPR as well as to paint a clearer picture of data privacy and best practices regarding privacy policies and their level of compliance with the GDPR. We leveraged the data mining and machine learning capabilities of our PrivacyCheck tool and three corpora totaling 550 privacy policies to compare privacy policies, pre- and post- GDPR. In addition, the UT CID research team manually answered 10 fundamental questions representing core tenets of the GDPR to judge a privacy policy's level of compliance to the GDPR.

The verdict is that, with modest changes, most of the privacy policies were able to satisfy many but not all GDPR requirements. The most notable non-compliance of the investigated privacy policies was found when policies fail to indicate compliance with a GDPR requirement, either affirmatively or negatively. Consequently, the most notable non-compliance with the GDPR results when an organization lacks transparency and explicit disclosure of their processing and protection of consumer personal information. If PII is the "currency of the Internet," these findings mark continued alarm about the agency of consumers to protect and secure their PII assets.

We found that websites have modestly changed their privacy policies after the GDPR. Changes have often been geared toward compliance. This is not surprising since the penalties for non-compliance far exceed prior regulations. Notably, privacy policies have increased their level of protection for two PrivacyCheck factors, PII of children and data aggregation, by 22% and 13%, respectively.

With respect to some privacy factors, most notably sharing with law enforcement and deleting/editing information, our research findings show that the protection level of the privacy policies

[8]California Consumer Privacy Act (CCPA): https://www.caprivacy.org.

has decreased by 8% and 13%, respectively. This observation is not surprising as the GDPR stresses the importance of sharing data with law enforcement. While many may find an increased sharing of a consumer's personal data with law enforcement without a subpoena to be alarming, an organization's increased willingness to share consumer PII with law enforcement also increases the organization's privacy policy compliance with the GDPR since the GDPR encourages such sharing. Investigating consumer rights around the globe, such as the right for consumers to edit/delete their information, our research shows a 13% decrease in protection and an alarming level (57%) of non-compliance to the GDPR in new policies after the release of the GDPR in May 2018. Consequently, the policies strip consumers of their agency and ability to take the necessary actions to protect, control, or ensure the accuracy of their personal information.

Overall, the landscape of privacy policies after the GDPR is promising when considering all 10 privacy factors measured by PrivacyCheck. For many factors, over 90% of the privacy policies we considered were in compliance with the GDPR, including the following factors: sharing PII with ads and marketing, collecting and sharing location, collecting and sharing PII of children, sharing with law enforcement, and data aggregation. It is important to note that the apparent decrease of protection with regard to sharing a consumer's PII with law enforcement discussed above does not create non-compliance with the GDPR, as the GDPR encourages such sharing in many instances.

Based on the manual examination of policies, when non-compliance does appear, it is often in failing to explicitly indicate compliance. We identify the following areas for improvement in privacy policies:

- Many policies fail to mention whether they encrypt data while at rest.
- Many do not mention if and when they notify the supervisory authority in case of a data breach.
- Some US-based policies protect children under 13 as per US regulation, and hence do not necessarily protect children between 13 and 16 years of age.
- Privacy policies lack a 100% consensus to require informed consent.

All in all, the GDPR has made progress in enforcing its data protection regulations. More work is necessary, particularly in the area of granting users the right to edit, update, and delete their data, to entirely fulfill the GDPR's promise.

Finally, we envision future PrivacyCheck advances with support for the new questions to further identify GDPR compliances in the context of the 10 GDPR tenets identified in this work. The corpora built in this work can serve as the training sets for PrivacyCheck extensions to include the new questions regarding GDPR compliance.

ACKNOWLEDGMENTS

We would like to thank the authors of Polisis for sharing their corpus of privacy policies. We thank R. Sean McCleskey, JD, for his legal analysis of the GDPR and identified tenets. We also thank the undergraduate students from the Spring 2019 EE379K Information Security and Privacy course at the University of Texas at Austin as well as Turan Z. Vural, David Liau, Kai Chih Chang, and Teng-Chieh Huang for their contribution to manually investigate privacy policies. Finally, we would like to thank Ryan Anderson for proofreading this manuscript and the anonymous reviewers for their comments.

REFERENCES

[1] Benjamin Andow, Samin Yaseer Mahmud, Wenyu Wang, Justin Whitaker, William Enck, Bradley Reaves, Kapil Singh, and Tao Xie. 2019. Policylint: Investigating internal privacy policy contradictions on Google play. In *28th USENIX Security Symposium (USENIX Security'19)*. 585–602.

[2] Internet Archive. [n.d.]. Way Back Machine. Retrieved August 24, 2019, from https://web.archive.org/.
[3] Giuseppe Contissa, Koen Docter, Francesca Lagioia, Marco Lippi, Hans W. Micklitz, Przemysław Pałka, Giovanni Sartor, and Paolo Torroni. 2018. Claudette meets GDPR: Automating the evaluation of privacy policies using artificial intelligence. *Available at SSRN 3208596* (2018).
[4] Martin Degeling, Christine Utz, Christopher Lentzsch, Henry Hosseini, Florian Schaub, and Thorsten Holz. 2019. We value your privacy... Now take some cookies. *Informatik Spektrum* 42, 5 (2019), 345–346.
[5] Kassem Fawaz, Thomas Linden, and Hamza Harkous. 2019. The applications of machine learning in privacy notice and choice. In *2019 11th International Conference on Communication Systems & Networks (COMSNETS'19)*. IEEE, 118–124.
[6] FTC. 2000. Privacy Online: Fair Information Practices in the Electronic Marketplace: A Federal Trade Commission Report to Congress. Retrieved October 21, 2015, from https://www.ftc.gov/reports/privacy-online-fair-information-practices-electronic-marketplace-federal-trade-commission.
[7] FTC. 2010. Exploring Privacy: An FTC Roundtable Discussion. Retrieved May 21, 2015, from https://www.ftc.gov/sites/default/files/documents/public_events/exploring-privacy-roundtable-series/privacyroundtable_march2010_transcript.pdf.
[8] FTC. 2012. Protecting Consumer Privacy in an Era of Rapid Change: Recommendations For Businesses and Policymakers. Retrieved May 21, 2015, from https://www.ftc.gov/reports/protecting-consumer-privacy-era-rapid-change-recommendations-businesses-policymakers.
[9] Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G. Shin, and Karl Aberer. 2018. Polisis: Automated analysis and presentation of privacy policies using deep learning. In *27th USENIX Security Symposium (USENIX Security'18)*. 531–548.
[10] Timothy Libert, Lucas Graves, and Rasmus Kleis Nielsen. 2018. Changes in third-party content on European news websites after GDPR. Retrieved Nov 30, 2020 from https://reutersinstitute.politics.ox.ac.uk/our-research/changes-third-party-content-european-news-websites-after-gdpr.
[11] Thomas Linden, Rishabh Khandelwal, Hamza Harkous, and Kassem Fawaz. 2020. The privacy policy landscape after the GDPR. *Proceedings on Privacy Enhancing Technologies* 2020, 1 (2020), 47–64.
[12] Marco Lippi, Przemysław Pałka, Giuseppe Contissa, Francesca Lagioia, Hans-Wolfgang Micklitz, Giovanni Sartor, and Paolo Torroni. 2019. CLAUDETTE: An automated detector of potentially unfair clauses in online terms of service. *Artificial Intelligence and Law* 27, 2 (2019), 117–139.
[13] Razieh Nokhbeh Zaeem and K. Suzanne Barber. 2017. A study of web privacy policies across industries. *Journal of Information Privacy and Security* 13, 4 (2017), 169–185.
[14] Rima Rana, Razieh Nokhbeh Zaeem, and K. Suzanne Barber. 2019. An assessment of blockchain identity solutions: Minimizing risk and liability of authentication. In *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI'19)*. IEEE, 26–33.
[15] Having Regard. 1980. Recommendation of the council concerning guidelines governing the protection of privacy and transborder flows of personal data. Retrieved Nov 30, 2020 from https://www.oecd.org/sti/ieconomy/2013-oecd-privacy-guidelines.pdf.
[16] Disconnect Me. 2014. Disconnect Me Privacy Icons. Retrieved March 15, 2016, from https://disconnect.me/icons.
[17] Norman Sadeh, Alessandro Acquisti, Travis D. Breaux, Lorrie Faith Cranor, Aleecia M. McDonalda, Joel R. Reidenbergb, Noah A. Smith, Fei Liu, N. Cameron Russellb, Florian Schaub, et al. 2013. *The Usable Privacy Policy Project*. Technical Report, CMU-ISR-13-119, Carnegie Mellon University.
[18] Welderufael B. Tesfay, Peter Hofmann, Toru Nakamura, Shinsaku Kiyomoto, and Jetzabel Serna. 2018. I read but don't agree: Privacy policy benchmarking using machine learning and the eu gdpr. In *Companion Proceedings of the the Web Conference 2018*. 163–166.
[19] Welderufael B. Tesfay, Peter Hofmann, Toru Nakamura, Shinsaku Kiyomoto, and Jetzabel Serna. 2018. PrivacyGuide: Towards an implementation of the EU GDPR on internet privacy policy evaluation. In *Proceedings of the 4th ACM International Workshop on Security and Privacy Analytics*. ACM, 15–21.
[20] ToS;DR. 2012. Terms of Service; Didn't Read. Retrieved March 4, 2015, from https://tosdr.org.
[21] European Union. [n.d.]. European Union Law. Retrieved August 24, 2019, from https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1566668063189&uri=CELEX:32016R0679.
[22] Shomir Wilson, Florian Schaub, Aswarth Dara, Sushain K. Cherivirala, Sebastian Zimmeck, Mads Schaarup Andersen, Pedro Giovanni Leon, Eduard Hovy, and Norman Sadeh. 2016. Demystifying privacy policies with language technologies: Progress and challenges. In *Proceedings of LREC 1st Workshop on Text Analytics for Cybersecurity and Online Safety*.
[23] Razieh Nokhbeh Zaeem, Suratna Budalakoti, K. Suzanne Barber, Muhibur Rasheed, and Chandrajit Bajaj. 2016. Predicting and explaining identity risk, exposure and cost using the ecosystem of identity attributes. In *2016 IEEE International Carnahan Conference on Security Technology (ICCST'16)*. IEEE, 1–8.

[24] Razieh Nokhbeh Zaeem, Rachel L. German, and K. Suzanne Barber. 2018. PrivacyCheck: Automatic summarization of privacy policies using data mining. *ACM Transactions on Internet Technology (TOIT)* 18, 4 (2018), 53.
[25] Razieh Nokhbeh Zaeem, Monisha Manoharan, and K. Suzanne Barber. 2016. Risk kit: Highlighting vulnerable identity assets for specific age groups. In *2016 European Intelligence and Security Informatics Conference (EISIC'16)*. IEEE, 32–38.
[26] Razieh Nokhbeh Zaeem, Monisha Manoharan, Yongpeng Yang, and K. Suzanne Barber. 2017. Modeling and analysis of identity threat behaviors through text mining of identity theft stories. *Computers & Security* 65 (2017), 50–63.
[27] Jim Zaiss, Razieh Nokhbeh Zaeem, and K. Suzanne Barber. 2019. Identity threat assessment and prediction. *Journal of Consumer Affairs* 53, 1 (2019), 58–70.
[28] Sebastian Zimmeck and Steven M. Bellovin. 2014. Privee: An architecture for automatically analyzing web privacy policies. In *23rd USENIX Security Symposium (USENIX Security'14)*. USENIX Association, San Diego, CA, 1–16. https://www.usenix.org/conference/usenixsecurity14/technical-sessions/presentation/zimmeck.

Received September 2019; revised March 2020; accepted March 2020