CompAι: A Tool for GDPR Completeness Checking of Privacy Policies using Artificial Intelligence

Orlando Amaral Cejas
orlando.amaralcejas@uni.lu
SnT, University of Luxembourg
Luxembourg

Sallam Abualhaija
sallam.abualhaija@uni.lu
SnT, University of Luxembourg
Luxembourg

Lionel Briand
lionel.briand@lero.ie
Lero SFI centre for Software Research
and University of Limerick,
School of EECS, University of Ottawa
Ireland, Canada

ABSTRACT

We introduce *CompAι* – a tool for checking the completeness of privacy policies against the general data protection regulation (GDPR). *CompAι* facilitates the analysis of privacy policies to check their compliance to GDPR requirements. Since privacy policies serve as an agreement between a software system and its prospective users, the policy must fully capture such requirements to ensure that collected personal data of individuals (or users) remains protected as specified by the GDPR. For a given privacy policy, *CompAι* semantically analyzes its textual content against a comprehensive conceptual model which captures all information types that might appear in any policy. Based on this analysis, alongside some input from the end user, *CompAι* can determine the potential incompleteness violations in the input policy with an accuracy of ≈96%. *CompAι* generates a detailed report that can be easily reviewed and validated by experts. The source code of *CompAι* is publicly available on https://figshare.com/articles/online_resource/CompAI/23676069, and a demo of the tool is available on https://youtu.be/zwa_tM3fXHU.

CCS CONCEPTS

• **Software and its engineering → Requirements analysis**; • **Security and privacy → Privacy protections**; • **Computing methodologies → Artificial intelligence**.

KEYWORDS

Requirements Engineering (RE), Regulatory Compliance, Privacy, the General Data Protection Regulation (GDPR), Artificial Intelligence (AI), Natural Language Processing (NLP), Machine Learning (ML).

ACM Reference Format:
Orlando Amaral Cejas, Sallam Abualhaija, and Lionel Briand. 2024. *CompAι*: A Tool for GDPR Completeness Checking of Privacy Policies using Artificial Intelligence. In *39th IEEE/ACM International Conference on Automated Software Engineering (ASE '24), October 27-November 1, 2024, Sacramento, CA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3691620.3695353

ASE '24, October 27-November 1, 2024, Sacramento, CA, USA

ACM ISBN 979-8-4007-1248-7/24/10.
https://doi.org/10.1145/3691620.3695353

1 INTRODUCTION

The reliance on digital services and huge concerns over privacy and data protection are growing side-by-side. Individuals learn about personal data collection and related processing activities through privacy policies that are typically associated with any software application. The General Data Protection Regulation (GDPR) in the European Union (EU) has been enforced since 2018 to harmonize the various privacy laws across Europe [5]. GDPR also affects organizations outside the EU as long as they process personal data of European residents. Software applications targeting the European market must issue GDPR-compliant privacy policies lest they are charged with hefty fines. According to GDPR, privacy policies should provide certain details about, e.g., which categories of personal data are collected, what is the source from which it is collected, for what purpose it will be processed, and with whom it will be shared.

In this paper we propose *CompAι*, a tool for **Comp**leteness checking of privacy policies using **A**rtificial **I**ntelligence. Given a privacy policy written in natural language (NL), *CompAι* examines its textual content against the privacy requirements stipulated in GDPR. Analyzing privacy policies has been extensively studied in the Requirements Engineering (RE) literature [3, 6–8]. Unlike existing approaches, *CompAι* was developed and empirically evaluated in close collaboration with legal experts from Linklaters LLP (a major law firm headquartered in London with a branch in Luxembourg). *CompAι* implements a set of rules in accordance with a comprehensive conceptual model that was proposed in our previous work [1, 4]. The model describes the exhaustive list of information types that can be present in any GDPR privacy policy. These information types can be categorized into mandatory types (directly imposed by GDPR) or optional types (following best practices). According to this categorization, the rules can produce violations when mandatory and optional information types are missing, respectively.

CompAι leverages natural language processing (NLP) and machine learning (ML) technologies. It further maintains trace links to the GDPR articles to help automatically provide appropriate explanations for each decision. Over a set of more than 200 privacy policies from various sources, *CompAι* has an average accuracy of ≈ 96% in detecting incompleteness violations, about 23% more accurate than a straightforward keyword-based approach. Since *CompAι* performs in-depth semantic analysis of the policy, its overall processing time varies according to the size of the input privacy policy. Time can reach up to few minutes for analyzing large policies. The main approach underlying *CompAι* has been proposed in our previous work [1, 4]. In this paper, we extend the implementation of *CompAı* to provide the functionalities through a web service with a user-friendly interface. Further, we present a detailed view of our tool implementation alongside a user study which we conducted with the legal experts from Linklaters regarding the tool's practical usefulness.

On the next page, Fig. 1 (top) illustrates a usage scenario on a demo privacy policy for an imaginary bank located in Japan, named *Hikari Bank Ltd.* To run *CompAı*, the end user must first answer a set of context-dependent questions. The answers are used to determine which rules need to be verified for checking the completeness of the input privacy policy. The end-user of *CompAı* can be legal experts aiming to categorize the content of a privacy policy for more efficient compliance analysis. Additionally, requirements engineers might use *CompAı* to verify the completeness of a given policy before eliciting privacy-related requirements. *CompAı* then fully parses and analyzes the semantics of the input policy. Finally, using the answers provided by the end user, together with the predefined rules, *CompAı* produces a final report which outlines for each rule whether it is applicable, and if it is satisfied. For each rule, satisfied or violated, *CompAı* further provides the explanation on what is expected according to the respective GDPR article. As the figure shows, the excerpt on *data subject rights*, i.e., the rights individuals have on their personal data, are missing the right to lodge a complaint which is then identified as a *violation* by the tool.

In the remainder of this tool demonstration paper, we first describe the tool architecture and illustrate its GUI. We further introduce a user study on the usefulness of the tool in practice.

2 TOOL ARCHITECTURE

CompAı can be accessed through this link https://compai.uni.lu/. Below, we present a walk-through of an end-to-end application of the tool on a demo privacy policy, named *Hikari* (introduced in Section 1). To analyze *Hikari*, *CompAı* performs six steps, depicted in Fig. 1 and explained next[1].

2.1 Upload Policy

This step uploads a privacy policy to be analyzed by *CompAı*. The tool accepts two formats, namely Microsoft (MS) Document and PDF. *CompAı* was originally developed for handling MS documents in our previous work [1, 4]. The PDF option involves simply transforming the privacy policy into an MS document. We support PDF in this web version to allow users to feed any privacy policy (typically available in PDF format) without having to do the format transformation manually. That said, we have no guarantee regarding the tool performance on PDF documents. Once the user selects the input privacy policy and clicks submit, the policy will be uploaded and the tool shows the next page, that is the questionnaire page.

2.2 Answer Questionnaire

As a prerequisite step to completeness checking, the user needs to answer six questions [1]. These questions capture details that depend on the context and are often left tacit in the privacy policies, such as the identity of the controller and, if applicable, the controller representative, the intention of data transfer, the collection source of personal data, and whether or not a data protection officer is required. The answers to the questionnaire contribute to validating whether GDPR compliance is at all relevant to the policy under analysis. For instance, if the controller who collects personal data is outside the EU, but its activities are carried out inside the EU, then the privacy policy is subject to GDPR compliance and the controller shall name a "controller representative" who resides in the EU. The provided answers further help select which rules are applicable in the completeness checking process. For example, if there is no intention in transferring personal data outside the EU, then the criterion related to that information type does not need to be checked. Next, the answers are passed on to step 5.

After submitting the answers, *CompAı* starts analyzing the uploaded privacy policy. This process might take up to few minutes depending on the size of the privacy policy.

2.3 Transform Text

In this step, *CompAı* builds semantic representations of the sentences in the input privacy policy. More details on this step can be found in our previous work [1].

2.4 Identify Information Types

In this step, *CompAı* predicts, for each sentence in the input policy, information types that are possibly present in that sentence. The example in Fig. 1 (on top) shows the identification of some data subject rights, namely the right to access, rectification, restriction, erasure, and object. The basis for this step is a comprehensive conceptual model that we created in close collaboration with legal experts from Linklaters. This prediction utilizes a combination of similarity-based, machine learning-based, and keyword-based approaches. The intermediary output, containing the identified information types, is passed on to the next step.

2.5 Check Completeness

This step takes user input (collected in step 2) and a set of completeness rules that we created according to the GDPR requirements. Using these two inputs, *CompAı* then checks the applicable rules based on the user's answers. The output of this step is then a detailed report with satisfied and violated rules. This output is passed on to the next step.

2.6 Generate Report

This step processes and edits the detailed report generated in the previous step to be more readable. *CompAı* presents to the end user the output at two levels. The first level consists of a summary page which describes the aggregated findings. In this page, the user can directly see the violations, warnings, as well as the GDPR requirements that are fulfilled in the input privacy policy. In addition, this page provides the reasoning behind each decision made, with the trace link to the relevant GDPR articles based on which the decisions are made. This way the output is explainable. The second level involves a more detailed view of the findings, i.e., the exact textual content that was identified by *CompAı* for each completeness rule. *CompAı* is user-friendly as it presents the details following a color

[1] Prior running the tool, one has to log in. To get the login credentials, send an email to sallam.abualhaija@uni.lu

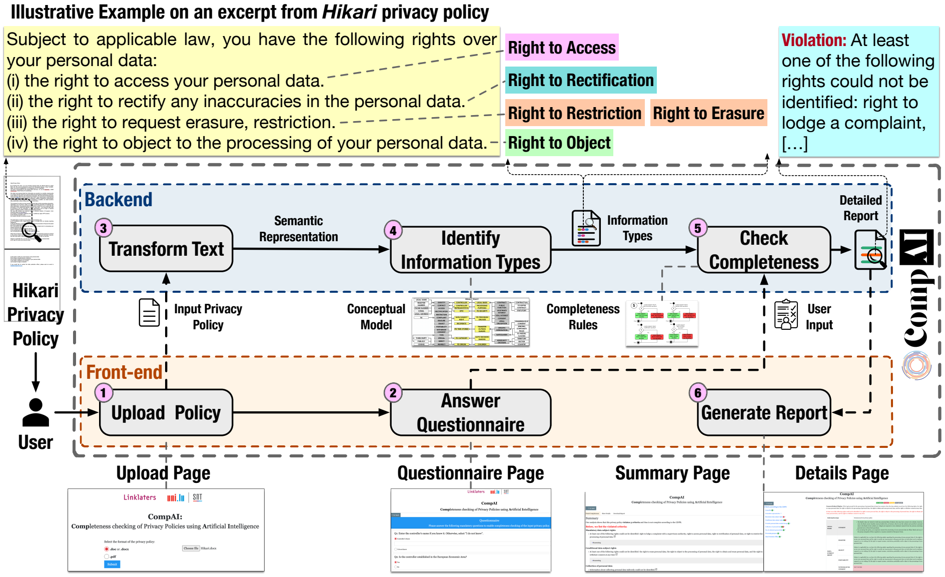

Illustrative Example on an excerpt from *Hikari* privacy policy
Subject to applicable law, you have the following rights over your personal data:
(i) the right to access your personal data.
(ii) the right to rectify any inaccuracies in the personal data.
(iii) the right to request erasure, restriction.
(iv) the right to object to the processing of your personal data.
Right to Access
Right to Rectification
Right to Restriction
Right to Erasure
Right to Object
Violation: At least one of the following rights could not be identified: right to lodge a complaint, [...]
Backend
3 Transform Text
Semantic Representation
4 Identify Information Types
Information Types
5 Check Completeness
Detailed Report
Input Privacy Policy
Conceptual Model
Completeness Rules
User Input
Hikari Privacy Policy
User
Front-end
1 Upload Policy
2 Answer Questionnaire
6 Generate Report
CompAI
Upload Page
Questionnaire Page
Summary Page
Details Page

Figure 1: *CompAɪ* Tool Architecture

coding: *green* represents the cases where no violation is identified, *red* represents violations, *orange* represents warnings, and *grey* represents the cases where the criterion is not applicable due to the selected combination of answers in the questionnaire page. Finally, the same details are also available as a downloadable report, automatically generated based on the findings. The report adheres to the color coding scheme. We further translate the legalese of GDPR requirements into plain English to improve readability, considering that the end users of *CompAɪ* are not necessarily legal experts.

3 USER STUDY

To evaluate the practical usefulness of *CompAɪ* in practice, we conducted a user study with two legal experts. The study material included two randomly-selected privacy policies ($\mathcal{P}_1$ and $\mathcal{P}_2$), automatically analyzed by *CompAɪ*. Table 1 lists the details of the analyzed privacy policies. For each policy, the table reports the total number of pages in that policy, the number of pages marked by *CompAɪ* as containing information types as well as the total number of information types identified by *CompAɪ*. The study was conducted over one session of ≈ two hours where the two legal experts and all the members of our research team participated. At the beginning of the session, we thoroughly explained the questionnaire of our user study to the legal experts. We then asked the experts to separately respond to all statements (provided below). We also asked the legal experts to verbalize their reasoning and discuss their rationale whenever they disagreed. To mitigate fatigue, we provided the study material to the experts one week in advance to familiarize themselves with the content.

Using Likert scales [2], we designed our study to collect feedback from the experts over four statements (**S1** – **S4**) and two follow-up statements (**S1-F** and **S2-F**), depicted in Fig. 2. Statements 1 and 2 are concerned with *false negatives* (i.e., missed information types) and *false positives* (i.e., falsely introduced information types), respectively. For statements **S1-F**, **S2-F**, **S3** and **S4**, the experts rated the questions on a five-point Likert scale. To account for pages that did not contain content relevant to our analysis, e.g., table of content, the experts were provided with an additional option "Not Relevant". The results from such pages were excluded from our analysis. The experts' feedback for statements **S1** – **S4** was collected on each page in $\mathcal{P}_1$ and $\mathcal{P}_2$.

Table 1: User study material details.

Privacy Policy	Pages	Pages containing $\mathcal{I}$	Number of $\mathcal{I}$
$\mathcal{P}_1$	10	10	66
$\mathcal{P}_2$	8	6	38
Summary	18	16	104

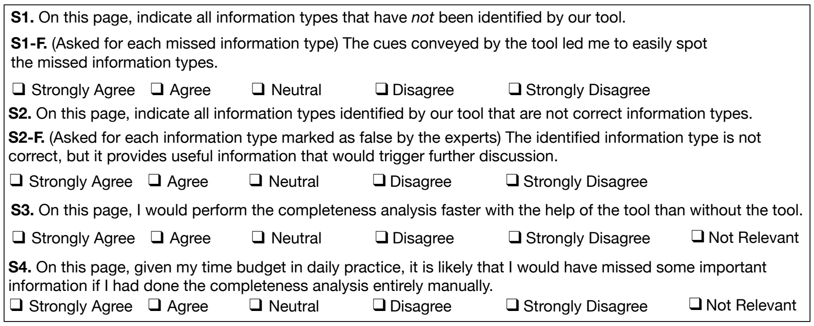
$\mathcal{I}$ *refers to information types.*

Table 2 summarizes the results of our user study. It reports for each policy the total number of information types found by *CompAɪ*, the number of information types marked as correct by the experts (true positives or TPs), the number of information types marked as incorrect by the experts (false positives or FPs), the number of information types missed by *CompAɪ* according to the experts' feedback (false negatives or FNs). We also report the precision (P) and recall (R) metrics, where P = TPs/(TPs+FPs) and R = TPs/(TPs+FNs).

Table 2: User study results.

Privacy policy	$\mathcal{I}$ found by *CompAɪ*	TPs	FPs	FNs	P(%)	R(%)
$\mathcal{P}_1$	66	62	4	5	93.9	92.5
$\mathcal{P}_2$	38	33	5	0	86.8	100
Summary	104	95	9	5	91.3	95.0

$\mathcal{I}$ *refers to information types.*



S1. On this page, indicate all information types that have *not* been identified by our tool.

S1-F. (Asked for each missed information type) The cues conveyed by the tool led me to easily spot the missed information types.

□ Strongly Agree □ Agree □ Neutral □ Disagree □ Strongly Disagree

S2. On this page, indicate all information types identified by our tool that are not correct information types.

S2-F. (Asked for each information type marked as false by the experts) The identified information type is not correct, but it provides useful information that would trigger further discussion.

□ Strongly Agree □ Agree □ Neutral □ Disagree □ Strongly Disagree

S3. On this page, I would perform the completeness analysis faster with the help of the tool than without the tool.

□ Strongly Agree □ Agree □ Neutral □ Disagree □ Strongly Disagree □ Not Relevant

S4. On this page, given my time budget in daily practice, it is likely that I would have missed some important information if I had done the completeness analysis entirely manually.

□ Strongly Agree □ Agree □ Neutral □ Disagree □ Strongly Disagree □ Not Relevant

Figure 2: User Study Questionnaire

With regard to statement **S1**, out of the 104 information types found by *CompAɪ*, the experts marked five FNs and 95 TPs. Consequently, the recall of *CompAɪ* on these privacy policies is 95%. For each FN, the experts answered the follow-up statement **S1-F**. Both experts provided a positive answer ("Strongly Agree") for all FNs, thus indicating that the output of *CompAɪ* can help in identifying FNs. With regard to **S2**, the experts marked as incorrect a total of nine information types (i.e., 9 FPs). Consequently, the average precision of *CompAɪ* on the analyzed privacy policies is 91.3%. For each FP, the experts were asked to provide an answer to the follow-up statement **S2-F**. The experts answered with "Agree" for six FPs and "Neutral" for the remaining three. This indicates that the text wrongly classified as containing information types can still point out some useful information regarding the completeness checking process.

With regard to **S3**, we collected a total of 16 (10 + 6) responses, one response for each page that contains information types. The experts provided nine answers as "Strongly Agree", and seven as "Agree". Overall, the experts agreed that *CompAɪ* helps them check the completeness of privacy policies more efficiently. With regard to **S4**, similar to **S3**, we collected a total of 16 responses. In this case, the experts responded with "Strongly Agree" 11 times, and "Agree" five times. These answers show that *CompAɪ* can indeed help the experts locate information types that they might have otherwise overlooked, given budget constraints.

In summary, our user study confirms the practical benefits of the automated support provided through our tool, *CompAɪ*. The results show that the output of *CompAɪ* can assist experts in the checking the completeness of privacy policies more efficiently. Instead of doing the analysis entirely manually, the experts can utilize the cues of the automated tool. We believe that the web version of our tool leads to additional advantages. Specifically, the interface and color coding scheme make applying the tool more user-friendly. The explainability feature also increases the credibility of the tool and further facilitates the validation of its findings.

4 CONCLUSION

We presented *CompAɪ*—a tool for checking the completeness of privacy policies against the general data protection regulation (GDPR). The current implementation of *CompAɪ* extends the solution presented in our previous work [1, 4] with a graphical user interface. *CompAɪ* combines natural language processing (NLP) and machine learning to check for the completeness of a given privacy policy according to GDPR provisions. For enabling the checking process, we created a conceptual model and a set of completeness rules through a qualitative study. Based on a user study we conducted with legal experts, *CompAɪ* has shown to be useful in practice to reduce the time required for manually analyzing privacy policies.

ACKNOWLEDGEMENTS

This research was supported by Linklaters and Luxembourg's National Research Fund (FNR) in whole, or in part, under grant references BRIDGES/19/IS/13759068/ARTAGO and NCER22/IS/16570468/NCER-FT. Lionel Briand was partly supported by the DG and CRC programmes of NSERC Canada. For the purpose of open access, and in fulfillment of the obligations arising from the grant agreement, the author has applied a Creative Commons Attribution 4.0 International (CC BY 4.0) license to any Author Accepted Manuscript version arising from this submission.

REFERENCES

[1] Orlando Amaral, Sallam Abualhaija, Damiano Torre, Mehrdad Sabetzadeh, and Lionel C. Briand. 2022. AI-Enabled Automation for Completeness Checking of Privacy Policies. *IEEE Transactions on Software Engineering* 48, 11 (2022), 4647–4674. https://doi.org/10.1109/TSE.2021.3124332

[2] Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology* (1932).

[3] Zeya Tan and Wei Song. 2023. PTPDroid: Detecting Violated User Privacy Disclosures to Third-Parties of Android Apps. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*. 473–485. https://doi.org/10.1109/ICSE48619.2023.00050

[4] Damiano Torre, Sallam Abualhaija, Mehrdad Sabetzadeh, Lionel C. Briand, Katrien Baetens, Peter Goes, and Sylvie Forastier. 2020. An AI-assisted Approach for Checking the Completeness of Privacy Policies Against GDPR. In *28th IEEE International Requirements Engineering Conference, RE 2020, Zurich, Switzerland, August 31 - September 4, 2020.*

[5] European Union. [n. d.]. General Data Protection Regulation. Accessed Nov. 07, 2021 [Online]. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L:2016:119:TOC

[6] Luca Verderame, Davide Caputo, Andrea Romdhana, and Alessio Merlo. 2020. On the (Un)Reliability of Privacy Policies in Android Apps. In *2020 International Joint Conference on Neural Networks (IJCNN)*. 1–9. https://doi.org/10.1109/IJCNN48605.2020.9206660

[7] Anhao Xiang, Weiping Pei, and Chuan Yue. 2023. PolicyChecker: Analyzing the GDPR Completeness of Mobile Apps' Privacy Policies. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*. Association for Computing Machinery, 3373–3387. https://doi.org/10.1145/3576915.3623067

[8] Le Yu, Xiapu Luo, Jiachi Chen, Hao Zhou, Tao Zhang, Henry Chang, and Hareton K. N. Leung. 2021. PPChecker: Towards Accessing the Trustworthiness of Android Apps' Privacy Policies. *IEEE Transactions on Software Engineering* 47, 2 (2021), 221–242. https://doi.org/10.1109/TSE.2018.2886875