Evaluating the Effect of Justification and Confidence Information on User Perception of a Privacy Policy Summarization Tool

Vanessa Bracamonte[1], Seira Hidano[1], Welderufael B. Tesfay[2] and Shinsaku Kiyomoto[1]

[1] *KDDI Research, Inc., Saitama, Japan*

[2] *Goethe University Frankfurt, Frankfurt, Germany*

Keywords: Privacy Policy, Automatic Summarization, Explanation, User Perception, Trust.

Abstract: Privacy policies are long and cumbersome for users to read. To support understanding of the information contained in privacy policies, automated analysis of textual data can be used to obtain a summary of their content, which can then be presented in a shorter, more usable format. However, these tools are not perfect and users indicate concern about the trustworthiness of their results. Although some of these tools provide information about their performance, the effect if this information has not been investigated. In order to address this, we conducted an experimental study to evaluate whether providing explanatory information such as result confidence and justification influences users's understanding of the privacy policy content and perception of the tool. The results suggest that presenting a justification of the results, in the form of a policy fragment, can increase intention to use the tool and improve perception of trustworthiness and usefulness. On the other hand, showing only a result confidence percentage did not improve perception of the tool, nor did it help to communicate the possibility of incorrect results. We discuss these results and their implications for the design of privacy policy summarization tools.

1 INTRODUCTION

The introduction of regulations such as the GDPR (European Parliament, 2016) has encouraged recent efforts to make privacy policies more understandable to users. However, despite serving as a defacto contractual document between the user and service provider, privacy policies remain too long and difficult for users to read and comprehend.

Alternatives to these lengthy pieces of text have been proposed, such as shorter notices in graphical and standardized formats (Gluck et al., 2016; Kelley et al., 2010) that can work to communicate information about the privacy policy to users. Since these formats are not employed by all companies, there are projects such as ToS;DR (ToSDR, 2019) which provides summaries of existing privacy policies. ToS;DR relies on a community of users to manually analyze and categorize the content of these privacy policies, which makes it very difficult to scale the work to cover every existing privacy policy.

There are also projects that propose to automatize the analysis of privacy policies using machine learning techniques. Privacy policy summarization tools are automated applications, implemented using different machine learning and natural language processing techniques, that analyze the content of a privacy policy text and provide a summary of the results of that analysis (Figure 1).

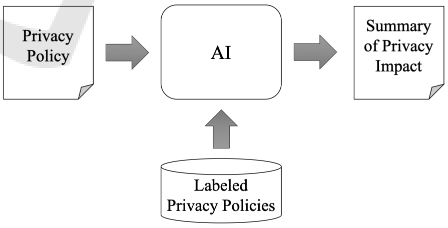


Figure 1: Automated privacy policy summarization.

Examples of these projects are Privee (Zimmeck and Bellovin, 2014), PrivacyCheck (Zaeem et al., 2018), Polisis (Harkous et al., 2018) and PrivacyGuide (Tesfay et al., 2018b). These tools can provide a solution to the problem of scale in the analysis of privacy policies, but they introduce a different challenge: users express concern regarding the trustworthiness and accuracy of a privacy policy summarization tool when they know that the process is automated (Bracamonte et al., 2019).

Bracamonte, V., Hidano, S., Tesfay, W. and Kiyomoto, S.
Evaluating the Effect of Justification and Confidence Information on User Perception of a Privacy Policy Summarization Tool.
DOI: 10.5220/0008986101420151
In *Proceedings of the 6th International Conference on Information Systems Security and Privacy (ICISSP 2020)*, pages 142-151
ISBN: 978-989-758-399-5; ISSN: 2184-4356

The design of the result summary of automated privacy policy summarization tools often follows guidelines for usable formats aimed at presenting privacy policy information. Research on alternative ways of presenting privacy policies indicates that shortened versions of these texts, supplemented with icons, can provide users with the necessary information for them to understand their content (Gluck et al., 2016). Not only the length of the privacy policy is important, but also how the information is presented: standardized graphical formats can better provide information than text (Kelley et al., 2009; Kelley et al., 2010). In addition to these design considerations, some privacy policy summarization tools also provide information related to the reliability and performance of the machine learning techniques used. However, there are no studies that evaluate how this information affects user perception of these type of tools. Although studies have evaluated perception and understanding of usable privacy policy formats produced by humans, aspects related to the reliability of the privacy policy information have naturally not been previously considered in the research.

The purpose of this study is to address this gap. To achieve this, we conducted an experiment to evaluate understanding and perception of the results of an automated privacy policy summarization tool. We created different conditions based on whether the results of the tool showed information about justification and confidence of the results, and asked participants about the content of the privacy policy and their perception of the tool in each of these conditions. The results show that justification information increased behavioral intention and trustworthiness and usefulness perception. Justification information also helped users qualify the answers provided by the tool, although the effect was not present for every aspect of the policy. Confidence information, on the other hand, did not have a positive effect on perception of the tool or on understanding of the results of the tool. We discuss these findings in the context of providing usable automated tools for privacy policy summarization and the challenges for the design of these tools.

2 RELATED WORK

2.1 Information Provided by Automated Privacy Policy Summaries

Automated privacy policy summarization tools mainly provide information about the privacy policy to the user: Privee (Zimmeck and Bellovin, 2014) and PrivacyCheck (Zaeem et al., 2018), for example, show a summary of a privacy policy based on pre-established categories and risk levels, and the visual design of the summary includes icons and standard descriptions. PrivacyGuide (Tesfay et al., 2018a) shows an icon-based result summary as well, and in addition provides the fragment of the original privacy policy. These tools are similar in the sense that they provide a standardized category-based summary of the privacy policy, although they use different criteria for that categorization and for assigning risk levels. Polisis (Harkous et al., 2018) takes a different and more complex approach for the privacy policy analysis and classification, but also provides standard categories and fragments of the original privacy policy text. A related chatbot tool, PriBot, returns fragments of the privacy policy in response to freely composed questions from users.

The automated tools mentioned sometimes provide explanatory information about the results. Two of the tools mentioned in the previous section, PrivacyGuide and PriBot, include information that may be considered as explanation of performance of the tool. PrivacyGuide provides a fragment of the original privacy policy that the tool identifies as related to a privacy aspect and uses to assign a risk level. PriBot, on the other hand, shows a confidence percentage that works as a proxy for how accurately the fragment it returns answers a user's question (Harkous et al., 2018).

2.2 Explanations of Automated Systems

When results are provided by automated tools, users have questions about the reliability of those results. One way of influencing this perception is through providing some explanation about the system. For example, offering some justification for outcomes can positively influence perception of accuracy (Biran and McKeown, 2017) and information about accuracy can improve trust (Lai and Tan, 2019). Although there is no predefined way of communicating to the user about the performance of automated privacy policy summarizing tools, existing tools provide some information. PrivacyGuide shows a fragment of the privacy policy, which can be classified as a justification or support explanation (Gregor and Benbasat, 1999). Similarly, information about the confidence of results such as the one provided by PriBot can also be considered a dimension of explanation (Wang et al., 2016). This serves as a measure of uncertainty (Diakopoulos, 2016) of the results and therefore as an indication of performance.

Explanation information, including confidence, has been found to positively influence trust when users interact with automation (Wang et al., 2016). However, the effect of explanations is not always positive. Research indicates that certain types of explanatory information about the performance of a system, for example an F-score accuracy measure, may not be useful in applications intended for a general audience (Kay et al., 2015). In addition, it has also been found that too much explanation could have a negative effect on aspects such as trust (Kizilcec, 2016). These studies show that simply providing more information may not result in a positive effect; therefore, it is important to evaluate the effect of explanatory information, such as justification and confidence, provided by privacy policy summarization tools.

One limitation in this area is that there are few user evaluations of automated privacy policy summarization tools. For Polisis, a user study was conducted that evaluated the perception of the accuracy of results; however, it was conducted independent from the interface (Harkous et al., 2018). The study found that users considered the results were relevant to the questions, although this perception differed from the measure of accuracy of the predictive model. A user study was also conducted to evaluate whether PrivacyGuide results, which found that the tool partially achieved the goal of informing users about the risk of a privacy policy and increasing interest in its content (Bracamonte et al., 2019). The study also found that users indicated concern about the trustworthiness of the tool and the accuracy of its results. However, these studies have not considered the effect of justification or confidence information shown by these tools, and how this explanatory information might affect trust.

3 METHODOLOGY

3.1 Experiment Design

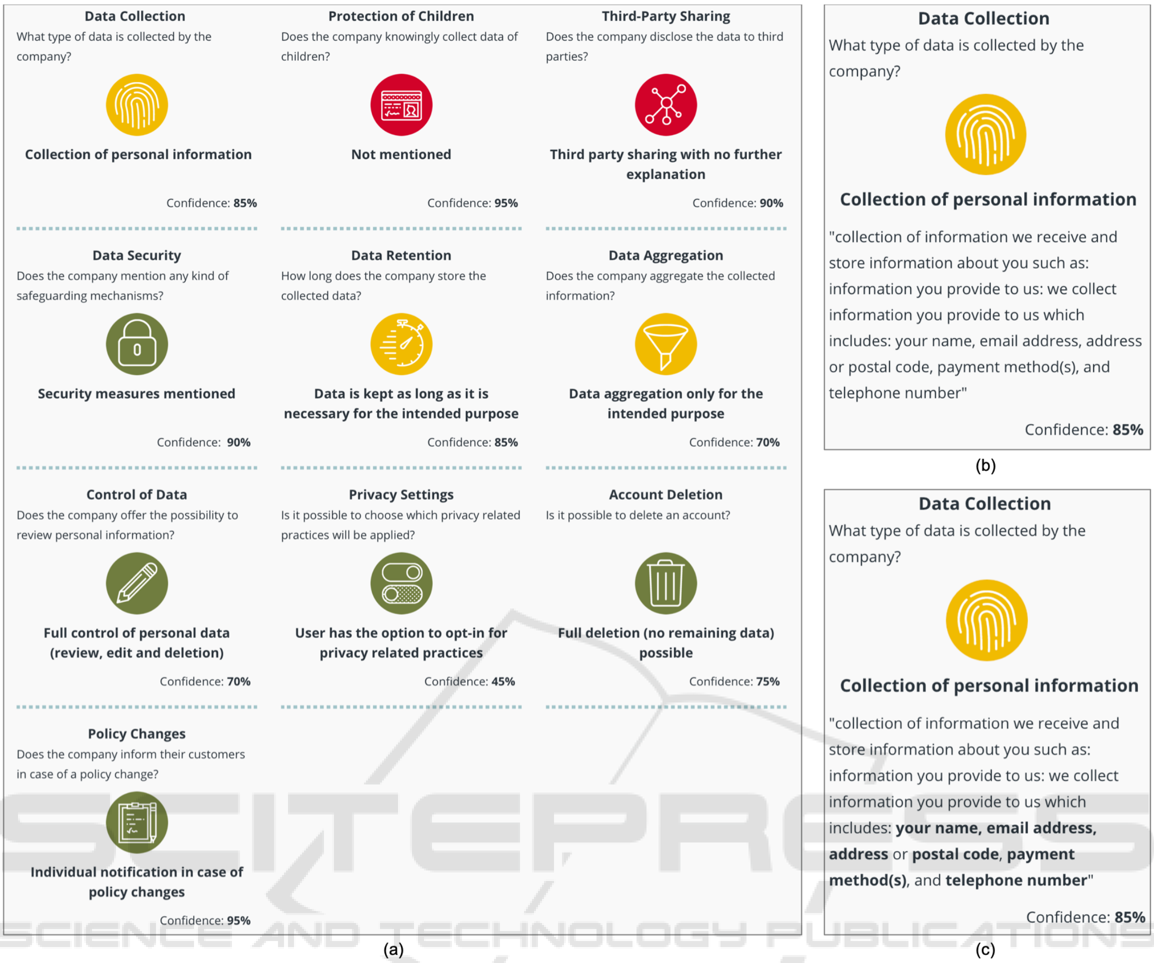
The experiment consisted of a task to view the results of the analysis of the privacy policy of an fictional online shop, and answer questions about the content of the privacy policy and about the perception of the tool in general. We used a between-subjects design, with a total of six experimental conditions. We defined the experimental conditions as follows. A Control condition that included only information about the result of the privacy policy summarization. A Confidence condition that included all the information from the Control condition and added a confidence percentage for the results. A Justification condition that included all the information from the Control condition and added justification in the form of a short fragments from the original privacy policy. A Highlight condition, which was a second form of justification where relevant words were emphasized in the privacy policy fragments. Finally, two conditions that showed both confidence percentage and justification (Justification + Confidence and Highlight + Confidence).

3.2 Privacy Policy Summary Result

We based the design of the privacy policy summary on PrivacyGuide and defined that the result would correspond to a low risk privacy policy, as defined by (Bracamonte et al., 2019). In PrivacyGuide, an icon in a color representing one of three levels of risk (Green, Yellow and Red) is assigned to each result category (privacy aspect) depending on the content of the privacy policy corresponding to that aspect (Tesfay et al., 2018b).

The Control condition result interface included icons and descriptions of the risk levels for each privacy aspect. The Justification condition result interface was based on the Control condition, and included in addition a text fragment for each privacy aspect. We selected the fragments from real privacy policies, by running PrivacyGuide on the English language privacy policies of well known international websites that also provided an equivalent Japanese language privacy policy. We chose those fragment results that matched the privacy aspect risk level we had defined, but took the fragment from the matching Japanese language privacy policy. This procedure resulted in fragments that were obtained from different privacy policies; therefore, we reviewed and modified the texts so that they would be congruent with each other in style and content. We also anonymized any reference to the original company. The Highlight condition result interface was based on the Justification result, and in addition emphasized the justification by highlighting words relevant to the corresponding privacy aspect.

The Confidence condition was based on the Control condition, and showed in addition a confidence percentage for each privacy aspect result. The confidence percentages were set manually. The Justification+Confidence and Highlight+Confidence conditions result interfaces showed all the information described previously for the Justified/Highlight and Confidence. Regarding the values of the confidence percentages, since confidence and justification information would be shown together for these last two conditions, we set the confidence values by manually evaluating how accurately the fragments represented the privacy aspect risk level. Confidence percentages for the privacy aspect results ranged from



Data Collection
What type of data is collected by the company?
Collection of personal information
Confidence: **85%**

Protection of Children
Does the company knowingly collect data of children?
Not mentioned
Confidence: **95%**

Third-Party Sharing
Does the company disclose the data to third parties?
Third party sharing with no further explanation
Confidence: **90%**

Data Security
Does the company mention any kind of safeguarding mechanisms?
Security measures mentioned
Confidence: **90%**

Data Retention
How long does the company store the collected data?
Data is kept as long as it is necessary for the intended purpose
Confidence: **85%**

Data Aggregation
Does the company aggregate the collected information?
Data aggregation only for the intended purpose
Confidence: **70%**

Control of Data
Does the company offer the possibility to review personal information?
Full control of personal data (review, edit and deletion)
Confidence: **70%**

Privacy Settings
Is it possible to choose which privacy related practices will be applied?
User has the option to opt-in for privacy related practices
Confidence: **45%**

Account Deletion
Is it possible to delete an account?
Full deletion (no remaining data) possible
Confidence: **75%**

Policy Changes
Does the company inform their customers in case of a policy change?
Individual notification in case of policy changes
Confidence: **95%**

(a)

Data Collection
What type of data is collected by the company?
Collection of personal information
"collection of information we receive and store information about you such as: information you provide to us: we collect information you provide to us which includes: your name, email address, address or postal code, payment method(s), and telephone number"
Confidence: **85%**

(b)

Data Collection
What type of data is collected by the company?
Collection of personal information
"collection of information we receive and store information about you such as: information you provide to us: we collect information you provide to us which includes: **your name, email address, address** or **postal code**, **payment method(s)**, and **telephone number**"
Confidence: **85%**

(c)

Figure 2: Experiment result screens. (a) Full results for the *Confidence* condition. (b) Fragment corresponding to one privacy aspect in the *Justification + Confidence* condition. (c) Fragment corresponding to one privacy aspect in the *Highlight + Confidence* condition. The result screens for the *Control, Justification and Highlight* conditions are similar to (a), (b) and (c), respectively, with the exception that the confidence percentage is not included.

70% to 95%, with the exception of Privacy Settings. For Privacy Settings, we set the confidence percentage to 45% and chose a fragment that did not accurately represent the corresponding privacy aspect. This was done to evaluate the influence of incorrect information and low confidence on users' response to questions about the content of the privacy policy. Figure 2 shows the details of the interface.

The interfaces also included a help section at the top of the result interface screen, which described every element of the results from the privacy aspect name to the confidence percentage (where applicable). Because users need time to familiarize themselves with elements in a privacy notice (Schaub et al., 2015), we included the help section to compensate for the lack of time, although such a section would not normally be prominently displayed.

3.3 Questionnaire

We included questions about the privacy policy, to evaluate participants' understanding of the privacy practices of the fictional company based on what was presented on the result interface. The questions were adapted from (Kelley et al., 2010) and addressed each of the privacy aspects (Table 1). We created the questions so that the correct option would be a positive answer (*Definitely yes* or *Possibly yes*) for all questions except for the question corresponding to the Protection of Children privacy aspect, where the correct option was left ambiguous. The only other exception was the Privacy Settings aspect, which was assigned low confidence percentage and an incorrect justification fragment (as defined in the previous section). Therefore, in the experimental conditions that included these pieces of information, we expected the

Table 1: Privacy policy content questions. Response options: Definitely yes, Possibly yes, Possibly no, Definitely no, It doesn't say in the result, It's unclear from the result.

Privacy aspect	Question
Data Collection	Does the online store (company) collect your personal information?
Privacy settings	Does the online store give you options to manage your privacy preferences?
Account deletion	Does the online store allow you to delete your account?
Protection of children	Does the online store knowingly collect information from children?
Data security	Does the online store have security measures to protect your personal information?
Third-party sharing	Does the online store share your personal information with third parties?
Data retention	Does the online store indicate how long they retain your data?
Data aggregation	Does the online store aggregate your personal information?
Control of data	Does the online store allow you to edit your information?
Policy changes	Does the online store inform you if they change their privacy policy?

Table 2: Questionnaire items. Response scale: Completely agree, Agree, Somewhat agree, Somewhat disagree, Disagree, Completely disagree.

Construct	Item
Useful	The application answers my questions about the privacy policy of the online store The application addresses my concerns about the privacy policy of the online store The application is useful to understand the privacy policy of the online store The application does not answer what I want to know about the privacy policy of the online store (Reverse worded)
Trustworthy	The results of the application are trustworthy The results of the application are reliable The results of the application are accurate
Understandable	The results of the application are understandable The reason for the results is understandable
Intention	I would use this application to analyze the privacy policy of various online stores I would use this application to decide whether or not to use various online stores
AI use	The use of AI is appropriate for this kind of application

correct answer to *not* be a positive answer.

We included items for measuring behavioral intention and perception of usefulness, understandability and trustworthiness of the tool (Table 2). The items were rated on a six-point Likert scale, ranging from Completely disagree to Completely agree. We also included a question addressing the perceived appropriateness of using AI for this use case.

3.3.1 Translation and Review

The questionnaire was developed in English; since we conducted the survey in Japan with Japanese participants, we translated the questionnaire with the following procedure. First, two native Japanese speakers independently translated the whole questionnaire, including the statements explaining the survey and privacy policy summarization tool. The translators and a person fluent in Japanese and English reviewed the translated statements one by one, verifying that both translations were equivalent to each other and had the same meaning as the original English statement.

The reviewers found no contradictions in meaning in this first step. The reviewers then chose the translated statements that more clearly communicated the meaning of the questions, instructions or explanations. Finally, the translators reviewed the whole questionnaire to standardize the language, since they had originally used different levels of formality.

3.4 Data Collection

We conducted the survey using an online survey company, which distributed an invitation to participate in the survey to their registered users. We targeted the recruitment process to obtain a sample with sex and age demographics similar to those of the Japanese population according to the 2101 census (Statistics Bureau, Ministry of Internal Affairs and Communications, 2010), but limited participation to users who were 18 years-old or older. Participants were compensated by the online survey company.

Participants were randomly assigned to one of the six experimental conditions and filled the survey online. We received the pseudonymized data from participants from the online survey company, which also included demographic data. In addition, the survey also registered the total time taken for the survey.

The survey was conducted from December 12-14, 2018.

3.5 Limitations

The study had the following limitations. In the study, we do not manipulate the risk level of the privacy policy, we only consider a privacy policy defined as low risk. The number of privacy aspects and corresponding risk levels result in a large number of possible combinations, making it impractical to test them all. Consequently, it may be that the results of the study are not generalizable to other risk levels besides the one chosen for the study.

In addition, we did not include a process to validate that the participants had indeed comprehended every aspect of the result interface, beyond the straightforward questions about the content of the privacy policy. We considered that if we included more detailed questions, the behavior of the participant would deviate further from a normal interaction with these type of tools. Nevertheless, this means that the results of this study reflect an evaluation of perception rather than objective measures of comprehension.

Lastly, in the study we used PrivacyGuide's privacy aspect categorization, which is based on the European Union's GDPR, and showed it to Japanese participants. However, we consider that the GDPR-based categories are relevant for our Japanese participants. For one, there is a degree of compatibility between the GDPR and Japanese privacy regulation (European Commission, 2019). And the Japanese language privacy policies used in the experiment come from international websites aimed at Japanese audiences, and are direct translations of English privacy policies created to comply with the GDRP.

4 ANALYSIS AND RESULTS

4.1 Data Cleanup

The online survey returned a total of 1054 responses. We first identified suspicious responses, defined as cases with no variability of extreme response (all questions answered with 1 or 6, which included reverse worded items for this purpose) and cases where the total survey answer time was lower than 125 seconds. We calculated this time considering a high reading speed and the number of characters in the online survey plus the result screen for the Control condition. With these criteria, we identified 110 cases which were manually reviewed and removed from further analysis.

Table 3: Sample Characteristics.

		n	%
Total		944	100%
Gender	Male	458	49%
	Female	486	51%
Age	19-20s	168	18%
	30s	175	19%
	40s	200	21%
	50s	185	20%
	60s	216	23%
Job	Government employee	35	4%
	Company Employee	373	40%
	Own business	59	6%
	Freelance	17	2%
	Full-time homemaker	174	18%
	Part time	112	12%
	Student	42	4%
	Other	27	3%
	Unemployed	105	11%

4.2 Sample Characteristics

The sample after data cleanup consisted of 944 cases (Table 3). 51% of the participants were female, and the age range was 19-69 years. The distribution these demographic characteristics is similar to that of the Japanese population (Statistics Bureau, Ministry of Internal Affairs and Communications, 2010), as established in the data collection process.

4.3 Effect on Responses to Questions about the Content of the Privacy Policy

We analyzed the categorical responses to the questions about the privacy policy using chi-square tests. We were interested in the differences between the responses to each privacy aspect question, so we used contingency tables to represent the relationship between questions and answers in each experimental condition. As indicated previously, we wanted to test whether participants had understood the results of the tool and whether differences in the information provided in each condition were reflected in their answers. The results of the chi-square test of independence are shown in Figure 3. Association

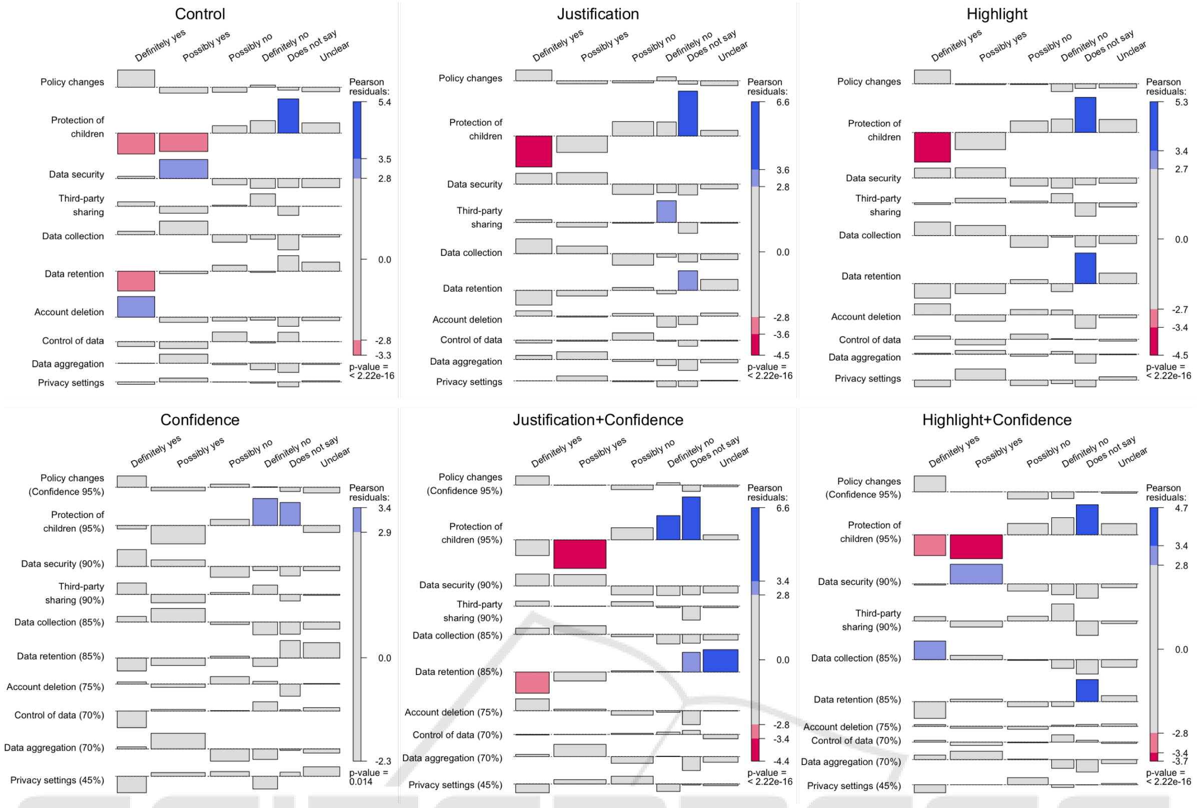


Figure 3: Association plot of the relationship between privacy policy questions and their responses for each experimental condition. Blue and red areas indicate significantly higher and lower response proportion than expected (i.e if all questions had the same response proportion), respectively.

plots (Meyer et al., 2006) were used to visualize areas with significantly higher or lower response proportion, compared between privacy aspect questions.

The responses in the Control condition provide a base for how participants understood the content of the privacy policy from the results of the tool. The majority of participants chose a positive answer for all of the questions except the one corresponding to Protection of Children, indicating that the result of the tool communicated the expected information in the Control condition. However, the results show that there were no differences in the proportion of responses corresponding to the question regarding Privacy Settings compared to other aspects in any of the experimental conditions. As can be observed in Figure 3, there are no significant differences in the proportions of responses to the Privacy Settings question compared to the responses to the Control of Data or Data Aggregation questions, for example. This lack of significant differences indicates that participants' responses were not influenced by either the low confidence percentage nor the incorrect justification in the result for the Privacy Settings aspect. However there is some evidence that at least some participants considered the justification information in their response. For the conditions that include justification, the proportion of *Does not say* responses in the Data Retention question is higher. A review of the fragment corresponding to this privacy aspect indicates that there is no mention of a specific time for the retention of the data. This lack of detail may have resulted in at least some participants considering that the question was not answered in the privacy policy. We do not see this difference in the Control condition nor in the Confidence condition, which do not include the justification fragment.

We also tested for differences in the time taken to answer the full survey between conditions. The distribution of time was similar for all conditions and highly skewed, so we used non-parametric Kruskal-Wallis tests for the difference in median. We found no significant differences, indicating that the additional information of justification and confidence conditions did not have an influence on the time taken to finish the survey.

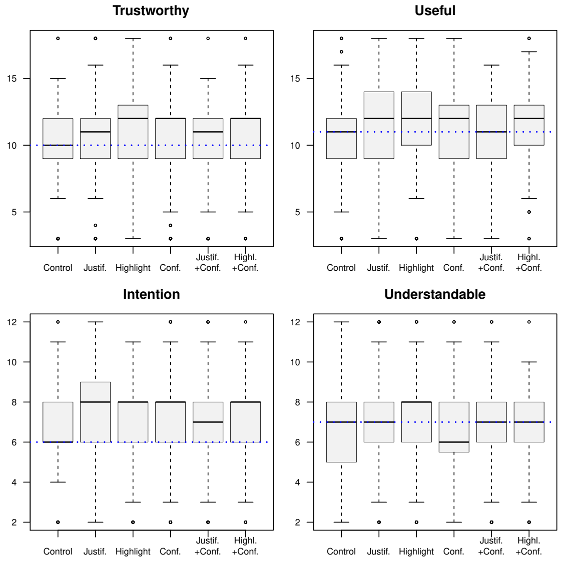


Figure 4: Box plots for all variables. The blue line indicates the median for the *Control* condition. The detail of significant differences according to Dunn's test are indicated in Table 5.

Table 4: Results of Kruskal-Wallis' test for the difference in median between conditions.

	Chi-squared (df=5)	p
Intention	11.694	0.033
Useful	17.217	0.004
Understandable	11.715	0.039
Trustworthy	17.028	0.004

Table 5: Results of Dunn's test for multiple comparisons. P-values for comparisons that were non-significant (p>0.5) are not shown. No significant differences were found for perceived understandability.

	Useful	Trustw.	Intention
Control - Justification		0.049	0.017
Control - Highlight	0.010	0.009	
Control - Highlight+Conf.	0.048	0.026	
Highlight - Confidence	0.035		

4.4 Effect on Perception of the Tool

We created composite variables by summing the items corresponding to attitudes and perception of usefulness, understandability and trustworthiness. We calculated a Cronbach's alpha measure for the items corresponding to each composite variable. In the case of usefulness, the recoded reverse-worded item was negatively correlated and we removed it from the analysis. After removal, Cronbach's alpha values indicated good internal consistency (all values above 0.9.) Figure 4 shows the median in each experimental condition for all variables.

All composite variables had a similar non-normal distribution shape; therefore, we used non-parametric Kruskal Wallis tests for the difference between their medians, and Dunn's test for multiple comparisons. To control for false positives, p-values were adjusted using the Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995). We found significant differences between groups for all variables, according to the results of Kruskal-Wallis tests (Table 4). We conducted post-hoc comparisons using Dunn's tests to evaluate which of the groups were significantly different. Table 5 shows the detailed results.

The Highlight and Highlight+Confidence conditions were more positively perceived in terms of usefulness and trustworthiness than the Control condition. For behavioral intention, we found a significant difference only between the Control and Justification conditions. In addition, we found no differences in the perception of any of the variables of interest between the Control and Confidence conditions. We also did not find significant differences between the Justification and Highlight conditions; we consider that this may be due to the relatively subtle effect of bolding the words. Nor were there significant differences between similar experimental conditions with or without confidence information. In addition, although the Kruskal-Wallis test indicated significant difference for understandability, the post hoc Dunn test did not find significant differences between any condition, based on the adjusted p-value.

Finally, with regard to the appropriateness of using AI for summarizing privacy policies, the results of a Kruskal-Wallis test showed that there were no significant differences between conditions. The median value was 4 (”Somewhat agree”) for all conditions except the Control condition, which had a median value of 3 (”Somewhat disagree”).

5 DISCUSSION

The results show that providing privacy policy information fragments as justification, with and without highlighted words, improved perception of the tool compared to not showing that information, albeit on different dimensions. On the other hand, the results show that confidence information did not have any influence. Based on previous research on the confidence explanations (Wang et al., 2016), we had expected that showing a confidence percentage would improve perception of trustworthiness in particular, but there was no effect on any of the measured perception variables.

The results show that the short summary format (Control condition) can inform users of the overall privacy policy contents as categorized by the tool. This result is in line with research on shorter privacy policy formats (Gluck et al., 2016). On the other hand, the additional explanatory information of justification and result confidence included in other experimental conditions did not greatly alter the responses to the questions about the privacy policy content, even when the fragment did not accurately justify the result or the confidence percentage was low. One possibility is that participants may have relied mainly on the privacy aspect's icons and descriptions to answer the questions about the privacy policy, even when there was additional information that was in contradiction of the overall result. This may have been due to the fact that we did not explicitly bring attention to the justification and confidence percentage, beyond including its description in the help section of the interface. In addition, the questions we asked participants were straightforward and for the most part targeted information that was already available in the elements of the Control condition interface.

However, the results indicate that participants did consider justification information, at least to some extent, as evidenced by the answers to the Data Retention question in conditions where the privacy policy fragment was shown. As for the incorrect fragment corresponding to the Privacy Settings aspect, it may be that participants were not sufficiently familiar with this type of settings, and therefore could not judge whether the fragment was incorrect or not. In the case of confidence information, another possibility is that the users did not think to question the results of the tool and therefore ignored the contradictory information of low confidence. This can happen when the user considers that the system is reliable (Wang et al., 2016). Figure 4 shows that the median of trustworthiness perception is higher than the midpoint for all conditions, which lends support to this hypothesis.

In general, the results of this study suggest that adding explanatory information in the form of justification can be beneficial to automated privacy policy summarization tools. However, post-hoc power analysis indicates that the sample is large enough to detect small differences, meaning that it is possible that statistically significant improvements in perception of usefulness and trustworthiness may not be relevant in practice. Conversely, the findings that not all explanatory information had a significant positive effect suggest that it is important to evaluate whether this additional information can truly benefit users, before considering adding it to the result interface of an automated privacy policy summarization tool. Although our results do not indicate that there would be negative effects if explanatory information is shown, future research should consider evaluating any possible tradeoffs in terms of usability. More research is needed to identify what information to present to users in order to improve efficacy as well as perception of these automated privacy tools.

6 CONCLUSIONS

In this paper, we conducted an experimental study to evaluate whether showing justification and confidence of the results of an automated privacy policy summarization tool influences user perception of the tool and whether this information can help users correctly interpret those results. The findings suggest that showing a privacy policy fragment as justification for the result can improve perception of usefulness and trustworthiness of the tool, and can improve intention of using the tool. On the other hand, information about the confidence of results did not appear to have much influence on user perception. Moreover, the findings also indicate that even when the confidence information indicated higher uncertainty of the results, the users did not rely on this information to interpret the results of the tool.

Altogether, the findings indicate that it may be worth considering adding explanatory information to help improve the perception of an automated privacy policy summarization tool, but that the type of explanation should be carefully chosen and evaluated, since explanatory information by itself may not be enough to help users understand the limitations of the results of the tool. Future research should investigate the type of explanatory information that these automated privacy tools should provide to users, as well as how to present that information in a usable way.

REFERENCES

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Biran, O. and McKeown, K. (2017). Human-centric Justification of Machine Learning Predictions. In *Proceedings of the 26th International Joint Conference on*

Artificial Intelligence, IJCAI'17, pages 1461–1467. AAAI Press.

Bracamonte, V., Hidano, S., Tesfay, W. B., and Kiyomoto, S. (2019). Evaluating Privacy Policy Summarization: An Experimental Study among Japanese Users. In *Proceedings of the 5th International Conference on Information Systems Security and Privacy - Volume 1: ICISSP,*, pages 370–377. INSTICC, SciTePress.

Diakopoulos, N. (2016). Accountability in Algorithmic Decision Making. *Commun. ACM*, 59(2):56–62.

European Commission (2019). European Commission - PRESS RELEASES - Press release - European Commission adopts adequacy decision on Japan, creating the world's largest area of safe data flows. http://europa.eu/rapid/press-release_IP-19-421_en.htm.

European Parliament (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46.

Gluck, J., Schaub, F., Friedman, A., Habib, H., Sadeh, N., Cranor, L. F., and Agarwal, Y. (2016). How Short Is Too Short? Implications of Length and Framing on the Effectiveness of Privacy Notices. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, pages 321–340. USENIX Association.

Gregor, S. and Benbasat, I. (1999). Explanations from Intelligent Systems: Theoretical Foundations and Implications for Practice. *MIS Quarterly*, 23(4):497–530.

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. USENIX Association.

Kay, M., Patel, S. N., and Kientz, J. A. (2015). How Good is 85%?: A Survey Tool to Connect Classifier Evaluation to Acceptability of Accuracy. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 347–356. ACM.

Kelley, P. G., Bresee, J., Cranor, L. F., and Reeder, R. W. (2009). A "Nutrition Label" for Privacy. In *Proceedings of the 5th Symposium on Usable Privacy and Security*, SOUPS '09, pages 4:1–4:12. ACM.

Kelley, P. G., Cesca, L., Bresee, J., and Cranor, L. F. (2010). Standardizing Privacy Notices: An Online Study of the Nutrition Label Approach. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1573–1582. ACM.

Kizilcec, R. F. (2016). How Much Information?: Effects of Transparency on Trust in an Algorithmic Interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 2390–2395. ACM.

Lai, V. and Tan, C. (2019). On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 29–38. ACM.

Meyer, D., Zeileis, A., and Hornik, K. (2006). The Strucplot Framework: Visualizing Multi-way Contingency Tables with **vcd**. *Journal of Statistical Software*, 17(3).

Schaub, F., Balebako, R., Durity, A. L., and Cranor, L. F. (2015). A design space for effective privacy notices. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 1–17.

Statistics Bureau, Ministry of Internal Affairs and Communications (2010). Population and Households of Japan 2010.

Tesfay, W. B., Hofmann, P., Nakamura, T., Kiyomoto, S., and Serna, J. (2018a). 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*, WWW '18, pages 163–166. International World Wide Web Conferences Steering Committee.

Tesfay, W. B., Hofmann, P., Nakamura, T., Kiyomoto, S., and Serna, J. (2018b). 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*, IWSPA '18, pages 15–21. ACM.

ToSDR (2019). Terms of Service; Didn't Read. https://tosdr.org/.

Wang, N., Pynadath, D. V., and Hill, S. G. (2016). Trust Calibration Within a Human-Robot Team: Comparing Automatically Generated Explanations. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*, HRI '16, pages 109–116. IEEE Press.

Zaeem, R. N., German, R. L., and Barber, K. S. (2018). PrivacyCheck: Automatic Summarization of Privacy Policies Using Data Mining. *ACM Trans. Internet Technol.*, 18(4):53:1–53:18.

Zimmeck, S. and Bellovin, S. M. (2014). Privee: An Architecture for Automatically Analyzing Web Privacy Policies. In *23rd USENIX Security Symposium (USENIX Security 2014)*, pages 1–16. USENIX Association.