Understanding the Impact of Data Privacy Regulations on Software and Its Stakeholders

Lucas J. Franke

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
in
Computer Science

Chris Brown, Chair
Eugenia Rho
Aaron Brantly

May 5, 2023
Blacksburg, Virginia

Keywords: GDPR, privacy, data, security, regulation

Understanding the Impact of Data Privacy Regulations on Software and Its Stakeholders

Lucas J. Franke

(ABSTRACT)

The General Data Protection Regulation (GDPR) is a comprehensive data privacy law that limits how businesses can collect personal information about their consumers living in the European Union. For our research, we aimed to evaluate the impact that the GDPR has on the open-source community, an online community that encourages open collaboration between software developers. We conducted a quantitative analysis of GitHub pull requests in which we compared pull requests explicitly related to the GDPR to other non-GDPR pull requests from the same projects. We also conducted a qualitative pilot study in which we interviewed software developers with experience implementing GDPR requirements in industry or in open-source. From our research, we found that GDPR-related pull requests had significantly more activity than other pull requests, but that open-source developers did not perceive a significant impact on their software development processes when implementing GDPR compliance. Industry developers, on the other hand, had a more negative outlook on the GDPR, and found implementation to be difficult. Our results indicate a need to involve software developers in the lawmaking process in order to create direct and realistic expectations for developers when implementing privacy policies.

Understanding the Impact of Data Privacy Regulations on Software and Its Stakeholders

Lucas J. Franke

(GENERAL AUDIENCE ABSTRACT)

The General Data Protection Regulation (GDPR) is a comprehensive data privacy law that limits how businesses can collect personal information about their consumers living in the European Union. For our research, we aimed to evaluate the impact that the GDPR has on the open-source community, an online community that encourages open collaboration between software developers. We conducted a quantitative analysis of GitHub, a major online open-source platform. We compared pull requests (major contributions to a project) explicitly related to the GDPR to other non-GDPR pull requests from the same projects. We also conducted a qualitative pilot study in which we interviewed software developers with experience implementing GDPR requirements in industry or in open-source. From our research, we found that GDPR-related pull requests had significantly more activity than other pull requests, but that open-source developers did not perceive a significant impact on their software development processes when implementing GDPR compliance. Industry developers, on the other hand, had a more negative outlook on the GDPR, and found implementation to be difficult. Our results indicate a need to involve software developers in the lawmaking process in order to create direct and realistic expectations for developers when implementing privacy policies.

Dedication

I dedicate my thesis to my big sister, Christina, for showing me that this was possible.

Acknowledgments

Thank you so much to my parents for supporting me throughout my college career. I could not have done this without you. Thank you to my sister, Cassandra, for helping me form a realizable plan of action for my research. Thank you to my advisor, Dr. Chris Brown, for guiding me through this entire process and believing in me. Thank you to Dr. Aaron Brantly and Dr. James Davis for providing valuable consultation during the conceptual phase of this project. Thank you to Huayu Liang for aiding me in collecting data and offering her expertise for this research. Thank you to Ben for listening to me patiently as I wrote my thesis, complained, and practiced my defense.

Contents

List of Tables

List of Abbreviations

GDPR General Data Protection Regulation

PII Personal Identifiable Information

PR Pull Request

The GDPR is a data privacy regulation in the European Union that controls how businesses may collect the personal information of their users.

PII is any information that can be used to identify an individual, such as their name, birth date, or address.

A PR is a major contribution of code to a GitHub project where developers can comment, collaborate, and contribute.

Chapter 1

Introduction

As technology advances, an increasing concern of the everyday consumer is that of data privacy. The "Big Four" tech conglomerates of Google, Facebook, Amazon, and Microsoft collectively store an estimated 1,200 petabytes of data (1.2 million terabytes) [20]. In the United States, user data is treated like a commodity to be bought and sold without the user's knowledge or explicit permission. All of this data collection and processing may used to direct the user experience with targeted advertisements as well as increase the time that users spend on social media apps. Most users are not aware of what data is being collected about them or how it is being used. However, over the last decade, this information has been gradually revealed in explosive increments, through fiery court cases such as the 2018 Facebook-Cambridge Analytica scandal, in which the personal information of an estimated 87 million users on Facebook was used by Cambridge Analytica to sell directly to the Trump presidential campaign [14]. Facebook users did not consent to their data being used to influence the presidential election, and were understandably outraged. With distrust in Big Tech building and concerns about the security of personal information becoming a common topic of conversation, European Union legislators were influenced to examine privacy protection more thoroughly and produce a groundbreaking new legislation: the General Data Protection Regulation (GDPR) [22].

The GDPR took effect in the European Union on May 25, 2018. It focused on the pro-

tection of user personal data to an unprecedented degree. At the time of the regulation's passing, no data privacy law of this scale had ever been created before, and even today, no federal law in the United States comes close to matching it. The key ideology surrounding the GDPR is "informed consent": before any data is collected about a user, the business must explain to the user in clear language what data will be collected, who it may be sold to, and how it will be used, and then obtain the user's explicit consent [3]. Users can withdraw consent, request an audit of their data, request changes to the data, and request their data be completely removed at any time, and the business must address the request in a timely manner [4]. This legislation gave people power over their personal information for the first time, and within the first year of the GDPR coming into effect, users made more than 144,000 complaints and queries to request information about and changes to their personal data [7].

While a majority of people agree that the GDPR is a huge step towards improving data security and user autonomy, such an unorthodox regulation is not without its many challenges. Google has been fined multiple times over the years just for their cookie consent policies: €100 million in 2020 for using tracking cookies without obtaining consent from their users, and €150 million on 2022 for deliberately hiding the option for users to opt-out of optional cookies [18]. Though businesses can face heavy penalties for non-compliance, the law itself is vague in some of its wording, leading to frustration in some software developers who struggle to understand how to properly comply with the law [5].

Despite its flaws, some states in the United States have started to adopt privacy laws after the GDPR. Most notably, the California Consumer Privacy Act, which came into effect in January 2020, was heavily inspired by the GDPR [11]. However, more research must still be done on the impact that the GDPR has had on software developers and the software

development process as a whole. Though we can see the GDPR's high-level impact on businesses through their policies and the results of judicial proceedings, as well as its impact on legislation through US regulations inspired by the law, we as researchers want to understand the role that government regulations play in the software development cycle, including the experiences of software developers who regularly implement these requirements. Through software developers, we can potentially learn why data leaks still occur despite the heavy hand of regulatory forces, and where the law has not taken into consideration the needs of developers. As we see more privacy laws appear in the United States at the state level, we can use the mistakes of previous regulations to guide future legislation and improve communication between the legislators who create the law and the software developers who have to implement it.

For our research, we focused on software development in the open-source community. Open source differs greatly from the Computer Science industry at large because it focuses on providing free software that anyone on the internet can contribute to, regardless of background or skill level. The community encourages developers to collaborate with each other and share ideas. At the same time, there is no guarantee that open-source software performs the way it is advertised. In our research, we found that on GitHub, a major open-source online platform, thousands of projects have been posted where developers had to implement GDPR compliance in some way. For such a strict government regulation where one can face severe repercussions for violating it, we wondered how open-source software developers navigate the law without always knowing the identity of their collaborators or the budget of big tech companies.

To guide our research objectives, we created a few research questions about the impact of the GDPR on open-source software developers:

1. **Which GDPR concepts are the most difficult for software engineers to implement?**
2. **How do software engineers' interactions with a legal team or lawyers affect the software development process?**
3. **What sentiments do software engineers hold about fulfilling GDPR requirements?**
4. **What impact does the GDPR have on open-source projects on GitHub that implement its requirements?**

To answer these questions, we conducted a mixed-methods empirical study: a quantitative analysis of projects on GitHub that implement GDPR compliance and a qualitative pilot study where we interviewed both industry and open-source software developers. In our quantitative analysis, we collected various metrics on pull requests (major code changes to a project) that directly involved discussion or implementation of the GDPR and compared them to other pull requests from the same GitHub project. In our qualitative study, we conducted semi-structured interviews where we asked developers about their sentiment about the GDPR, what concepts they found challenging to implement, and their experience working with or without legal aid to validate GDPR compliance.

Chapter 2

Background

2.1 General Data Protection Regulation

The GDPR was first proposed on January 25, 2012. On April 8, 2016, the regulation was adopted by the Council of the European Union, but its provisions did not become enforceable until two years later, in order to give EU Member States and businesses time to prepare for implementation. On May 25, 2018, the GDPR went into effect [1]. Since then, over $9 billion have been spent by US and UK companies in order to implement compliance [25]. Even still, hundreds of fines and penalties have been imposed upon data controllers and processors for non-compliance. The largest fine ever imposed was against Amazon Europe Core S.à.r.l. on July 16, 2021 for 746 million euro [2].

There are several key terms that define the entities that interact with personal identifiable information (PII): data subjects, data controllers, and data processors. Data subjects are the individuals whose personal data is collected. Data controllers are any entity - organization, individual, or otherwise - that owns, controls, and is responsible for the personal data. Data processors are any entity that processes the data for the data controller, like an accounting firm [16].

The GDPR[1] is split into eleven chapters and 99 articles. It grants a slew of powerful rights to data subjects to give them control over their PII. It also provides guidelines and requirements to data controllers and processors to understand how to properly handle PII.

The GDPR only applies to activities that fall within the scope of EU law. In general, this means it covers the collection and processing of the personal data of European Union citizens, regardless of whether or not the data processor or controller in question is based in the European Union. Therefore, US businesses are still required to implement GDPR compliance as long as they collect or process the the personal data of EU users.

Through the GDPR, data subjects are granted many unique rights that most data subjects residing outside the EU do not get to have. Overall, the data controller must be completely transparent with the data subject about what personal data is collected, why it is collected, how long the data is stored, and what data processors will have access to the data. The data subject has the right to access the personal data that has been collected about them. They also have the right to promptly rectify any incorrect data about them. The data subject also has the right to the erasure of their personal data "without undue delay", which means that upon request, the data must be deleted as soon as it is no longer needed for its intended purpose. The data subject also has the right to restrict the processing of their data by the data controller, where the data is no processed but must still be saved. The data subject also has the right to data portability, meaning that they may receive from the controller all of their personal data in a "machine-readable" format, as well as transfer that data to another controller as they see fit. The subject may object to the use of their data for any reason, whereupon the controller must provide a legitimate reason to continue processing the data that supersedes any relevant rights or freedoms of the subject. Data subjects also have unique

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

rights to restrict the use of their personal data in automated processing. However, all these rights may be conditionally revoked by the state under certain circumstances, which include: if national or public security is at risk, if the rights of other individuals may be jeopardized, or if such data processing is necessary for the investigation and prosecution of criminal offenses.

Data controllers and processors may suffer harsh penalties for violating data subjects' rights. Fines can reach as high as €200 million or "up to 4 % of the total worldwide annual turnover of the preceding financial year, whichever is higher" [6].

2.2 GitHub

GitHub[2] is an online platform for individuals and organizations to host their open-source software projects. Open-source software is freely available for anyone to use, distribute, modify, and build upon. Thus, GitHub offers an extremely collaborative atmosphere for software developers to share their code online and contribute on others' projects. However, with the open nature of GitHub, there is also very limited liability for the developers that post code on the website. This means that all code on GitHub is meant to be used "at your own risk", as there is no guarantee that the developer has posted code that is secure or even runs correctly. Where our interest as researchers lies is in how software developers navigate the open-source field as it pertains to the GDPR, a binding legal regulation with severe penalties and fines associated if the developer fails to fully comply with the stipulations encoded therein. If open-source software is offered freely but with no guarantees of accuracy, how do software developers feel comfortable using or posting open-source code that claims to implement GDPR compliance?

[2]https://github.com

There are several important keywords to understand the structure of projects hosted on GitHub. The projects themselves are called repositories. According to the creator's permissions, potentially anyone can contribute to a repository. The repository is divided into branches, with the main branch being the definitive version of the project. To contribute, users make new branches, which are copies of the current version of the repository that they can edit freely without affecting the main branch. In order to discuss plans for future development in the repository, users can create issues, which are simply text posts where the user can discuss problems with the current implementation, ideas for additional features, or solutions to existing problems. The issues can be assigned to users, inviting them to collaborate on that particular problem. When a user wishes to merge the changes they made on their branch onto the main branch (or any other branch), they create a pull request. In general, *pull requests* (PRs) are the final phase of the process of adding new code to a repository. GitHub automatically checks whether or not there are any conflicts in code between the main branch and the merging branch. If there are, the user must resolve them manually before the branch can be merged. The project owner may also include any number of automatic security checks to be run on the pull request code before merging. Depending on the repository settings, the code may also need to be manually reviewed by a certain number of other privileged users and approved. Those users can request changes to the code as necessary, whether it be a request for additional features to be added, or criticisms on the current implementation. There are three possible states of a pull request: open, closed, and merged. If a pull request is open, it can still be interacted with and remains available as a potential contribution to the main branch. When a pull request is closed, it is no longer interactable, and the code is effectively no longer in consideration to be merged. When a pull request is merged, the code has been added to the main branch and accepted as a valuable contribution.

In order to collect data from GitHub, we made use of GitHub's REST API[3]. API stands for application programming interface, and it defines a set of tools through which individuals can interact with an application. A REST API is a type of API that conforms to the representation state transfer architectural style, or REST. Through the GitHub API, it is possible to search for repositories, issues, pull requests, and more that meet a user-defined set of criteria. From these searches, we can collect raw data about the search results, such as the number of commits, comments, and collaborators.

[3]https://docs.github.com/en/rest?apiVersion=2022-11-28#authentication

Chapter 3

Review of Literature

3.1 GDPR in Open Source

One notable paper [21] created a tool called GDPRbench that is used to assess the GDPR compliance of databases. The tool was based off of the Yahoo! Cloud Serving Benchmark (YCSB) which measures the performance and efficiency of databases. The researchers modified the tool so that in addition to measuring performance, it also scores the database based on a variety of privacy and security factors. The researchers developed and released their tool on GitHub, allowing anyone to make use of their research. We found this paper to be a valuable indication of reputable GDPR work being done in the open-source field. This paper was one of the driving forces that directed our interest to the open-source community as the focus of our research.

3.2 Software Engineering Strategies for Compliance

A common research goal among several papers was to explore strategies to aid organizations in implementing GDPR compliance. One paper [17] discusses a method called "continuous integration" and how it was adapted into a particular organization's approach to compliance. The paper also discussed the organization's various challenges with the GDPR along the way and the researchers' work in developing a tool to mitigate those concerns. This research deals

with generating software engineering strategies for industry developers to apply to their team and guide them through GDPR implementation. However, we do not expect open-source projects to follow the same structures and development patterns as those in industry, due the vast variability between the kinds of projects hosted on GitHub and the kinds of developers who contribute to them. This is part of the reason why we are so interested in open-source. It is a self-sufficient community but is fundamentally distinct from the tech industry, and as such, we are interested in how perspectives and implementation strategies differ as well.

Another paper [9] surveyed 56 different academic publications, analyzing different privacy implementation approaches over the years in order to create a taxonomy to definitively classify them. The paper hypothesized that this taxonomy could be used to guide future GDPR implementation. This paper focused on the functional aspects of implementing the law, while our research focuses on the perception and impact the law has had on software developers.

GDPR implementation differs according to the platform that the developer chooses. One paper [19] proposed a unique approach using blockchain technology to validate PII protection. The Blockchain is an emerging technology that still has many flaws with data protection, especially due to its unregulated nature. This paper offers novel solutions to issues that users face when making blockchain transactions. Of course, our research focuses on a different field, open source, but, interestingly, we can draw similarities between the issues blockchain users face and those of open-source developers. Being unregulated, the Blockchain can be just as unpredictable and unreliable as contributions between developers in the open-source community - a property worth investigating further when strict regulations like the GDPR come into play.

With how complex the interpretation of the GDPR can be, other researchers have attempted

to automate GDPR compliance [13]. This paper describes the creation of a scalable tool to automatically verify compliance. The tool translates GDPR requirements into functional code that the developer can use for their own project. This kind of tool is something we would like to apply our research towards in future work. The information we have gathered could be used to generate tools to ease the burdens of developers.

3.3 Interviews

Research has been done in the past to interview software developers about their perspective on the GDPR. In one paper [10], researchers interviewed software developers across the spectrum from novice to experienced, and found that some of the greatest barriers for these developers to adopt GDPR principles were lack of familiarity, lack of precedented techniques for implementation, lack of helpful online resources, and the lack of prioritization from their employer or clients. The paper also found that developers generally do not prioritize privacy features in their projects, focusing instead on functional requirements. Similarly, another paper [12] interviewed senior engineers involved in research and industry to understand their motives and struggles when implementing privacy compliance. This paper found that developers struggle with legal interactions, and perceive a lack of autonomy and control through their attempts at compliance. While these papers take similar approaches to our research, ultimately our goals and questions are distinct, since we are specifically interested in the perspective of open-source developers.

Another paper [23] also conducted interviews; however, these interviews were at the organizational level. The paper asked organizations, big and small, about how feasible they believed the requirements in the GDPR were to implement. Big organizations in general

were the most confident about their ability to comply with the GDPR, as well as smaller organizations who specialized in security. Other smaller organizations struggled with compliance due to the overall breadth and ambiguity of the GDPR, especially in regards to the qualitative requirements. The requirements surrounding database transparency also posed a challenge to the small organizations, as they found it difficult to map out their complex data structures. This paper was published in 2018, right when the GDPR was going into effect, so these interviews were mainly to gather the perceptions and expectations of these organizations about their plans to incorporate GDPR regulations into their products. With time, it is certainly possible and almost expected that these perceptions will shift as the organizations adapt to reality. Our study was conducted in late 2022 through early 2023, so enough time has passed for opinions to settle on a foundation of real-world implementation and precedent. Our interest is also less focused on the organizational level and rather on the software engineers who implement GDPR compliance in open source.

Rather than interviewing software developers or organizations, another paper [15] surveyed internet users. The researchers compared privacy protection goals that developers follow to the privacy features that consumers actually value. As a result, the paper found a clear disconnect between websites' privacy policies and the needs of users. This research aimed to suggest to developers a set of privacy policies that may be more beneficial to users according to the users themselves. In this way, the paper follows the same user-centric design philosophy as the GDPR. It differs from our research since our research aims to investigate the impact of privacy regulations on developers, while this research focuses on consumers.

Historically, privacy practices and discussions started long before the GDPR went into effect. One paper [24] catalogues the evolution of privacy practices as they transitioned from guidelines self-imposed by developers to expansive legislation created by regulators and legal

authorities. The paper also discusses an organization called Project CANDID funded by the EU to consult with experts in the fields of law, software development, and entrepreneurial pursuits, as well as the perspectives of civil rights associations and ordinary users. A large portion of this paper focuses on the discussions between individuals involved within Project CANDID, similar to our interest in developer perspectives in our own research. However, this paper trends towards discussions of smart cities and infrastructures, with no discussion of the open-source community at all.

Chapter 4

Methodology

Before doing any large-scale automation of our data collection, we spent time manually examining the results we were able to acquire through GitHub's search engine by simply searching for content associated with the GDPR. GitHub sorts its search results into different sections based on the type of contribution it is, whether it be a pull request, issue, discussion, or the entire repository. While examining these search results, we took notes on a number of important details expressed within: we looked for what part of the GDPR exactly the developers were trying to comply to, whether we believed that the developer's solution actually complied with the GDPR or not, suggestions provided by other developers on the issue in question, and commentary provided by developers on their subjective opinion about the GDPR. These initial examinations helped us to form our research questions and interview questions, as well as qualitatively assess the value of using the GitHub Search API to collect information on projects related to the GDPR.

4.1 GitHub Analysis

For our quantitative part of our study, we collected relevant information about pull requests on GitHub. We used GitHub's REST API to search for all GitHub pull requests on the platform that mentioned the GDPR explicitly in either the title, labels, code, or comments. We chose pull requests because pull requests represent actual code under review to be merged

or has already been merged into the repository. We found this to be the best representation of real contributions towards GDPR-related projects. We also chose to search for pull requests that specifically mention the GDPR in writing in order to be absolutely assured that the projects worked on GDPR-specific implementation, rather than some other legal or security requirement. We collected a sample of 998 GDPR-related pull requests. From those pull requests, we saved the following information about them: the number of comments, the creation date, the date of the latest update, the amount of time the pull request remained active, when, if at all, the pull request was closed, the number of commits, the number of lines of code that would be added to the main branch (additions), the number of lines of code that would be removed from the main branch (deletions), the number of changed files, and whether or not the pull request was currently merged, closed, or open. Then, from those search results, we retrieved information on the repositories where those pull requests were posted. We collected all other pull requests on those repositories that did not explicitly mention the GDPR. By doing this, we ended up with a sample size of 12,217 non-GDPR-related pull requests. We retrieved the same parameters as the GDPR-related pull requests and compared them with each other.

4.2 Pilot Interviews

From the GDPR-related pull requests we collected for our quantitative analysis, we collected a list of all GitHub users who collaborated on the first 98 pull requests. We used this list to determine individuals to reach out to in order to conduct an interview. We were only able to contact users who provided some form of contact information in their GitHub profile. We contacted 60 GitHub users via email and received 2 responses from users interested in participating in an interview. For each interview, we asked the individual several open-ended

questions related to the GDPR:

1. *(RQ 3)* What meaningful impact, if any, do you believe the GDPR has had on data security and privacy?

2. *(RQ 3)* How do data protection laws like the GDPR affect market incentives to provide secure software?

3. *(RQ 1)* What GDPR concepts do you find the most difficult or frustrating to implement?

4. *(RQ 2)* During your software development projects, do you frequently consult with a legal team, and if so, how does this impact the development processes?

5. *(RQ 2)* Have you had to specifically seek out legal consultation on GDPR-related issues, and if so, how did that affect your development process? If not, how did you assess GDPR compliance for your software projects?

These questions were intended to provide a loose guide through which the interviewer and interviewee could have a natural discussion about the interviewee's experience and perception of the GDPR. With consent, we recorded the conversations and saved the transcripts of the interview.

Aside from the participants we contacted through GitHub, we conducted an initial interview with an individual who had industry experience with implementing GDPR compliance, but did not have experience with open-source software. This individual's experience was a useful comparison point for the interviews we would later conduct with open-source developers.

Chapter 5

Results

5.1 Github Searching

From the GitHub REST API, we collected two sets of data: one dataset of pull requests that implemented some form of GDPR related material, and another dataset containing all other pull requests from the same repositories as the GDPR PRs. Our hypothesis was that the GDPR-related pull requests would have a longer active time, more commits, more additions, more deletions, more changed files, and have fewer cases of PRs being successfully merged. We used both the Mann-Whitney U test[1] and the t-test[2] in order to evaluate the statistical significance between the two datasets. In order for a parameter to be statistically significant, we were aiming for p-values smaller than 0.05. To evaluate the merged, closed, and open status, we conducted a t-test on each category to determine if there is a statistically significant difference between the distribution of PRs of each category within each dataset. We found that there was no statistically significant difference between the number of merged, closed, and open PRs in the GDPR dataset and the non-GDPR dataset. For the rest of the parameters, we conducted a Mann-Whitney U test for each parameter to determine if there is a statistically significant difference between the values in the GDPR-related dataset compared to the non-GDPR dataset. From this test, we found that the GDPR-related PRs had significantly more commits, comments, and additions, as well as having a longer

[1] https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_nonparametric/bs704_nonparametric4.html
[2] https://www.scribbr.com/statistics/t-test/

active time. There was a not a significant difference in the number of deletions. Finally, GDPR-related PRs had significantly fewer changed files than non-GDPR PRs.

Table 5.1: GDPR vs. Non-GDPR Development Activity Metrics

Characteristic	**Dataset**	**Mean**	**Median**	***p-value***
comments*	non-GDPR	0.373	0	-
	GDPR	0.758	0	$p < 0.0001$
active_time (seconds)*	non-GDPR	$3.23 * 10^7$	$2.68 * 10^7$	-
	GDPR	$5.28 * 10^7$	$5.29 * 10^7$	$p < 0.0001$
commits*	non-GDPR	2.04	1	-
	GDPR	2.78	1	$p < 0.0001$
additions*	non-GDPR	554.8	19	-
	GDPR	398.43	76	$p < 0.0001$
deletions	non-GDPR	290.9	11	-
	GDPR	377.2	14	$p = 0.285$
changed_files*	non-GDPR	4.32	1	$p = 0.000622$
	GDPR	9.58	1	-
merged	non-GDPR	0	0	-
	GDPR	0	0	$p = 0.841$
closed	non-GDPR	0	0	-
	GDPR	0	0	$p = 0.180$
closed_without_merging	non-GDPR	-	-	-
	GDPR	-	-	$p = 0.159$
open	non-GDPR	1	1	-
	GDPR	0.999	1	$p = 0.820$

*** denotes statistically significant results (**p-value < 0.05**)

5.2 Interviews

Our first participant (P1) was an individual with very minimal open-source experience. However, they had a large amount of experience implementing GDPR compliance as part of their industry job. Our second participant (P2) was a high-school student with four years of experience implementing GDPR compliance in open source. Lastly, our third participant (P3) had over 20 years of both industry and open-source experience.

All three participants held a generally positive opinion of the GDPR. All believed that the GDPR is a step in the right direction for granting users more control over their personal information. However, all three participants admitted to having limited knowledge about the actual content within the law, being only able to recite general ideas and principles of it.

P1 had the most negative outlook of the law overall. P1 cited issues with the ambiguity of the law. They found it difficult to understand how to be compliant, and mentioned how several judicial rulings in the EU effectively changed the interpretation of the GDPR as they were made, making it difficult to be sure that compliant software will remain compliant for a long period of time. When asked about a specific aspect of the GDPR they found difficult to implement, they stated that embedding content from one website into their website causes many compliance issues. In particular, they cited issues with embedding YouTube videos into their website, where to be compliant, they must ask the website visitor to consent to loading the video before it can appear on their screen. This in addition to qualms with how cookie consent policies are implemented lead P1 to believe that GDPR implementation leads to an overall negative user experience when visiting GDPR-compliant websites.

In order to ensure compliance, P1 consulted with legal experts regularly throughout the software development process. They stated that most of their GDPR implementation is actually initiated by direct requests from legal. P1 admitted to not always knowing what requirements were specifically GDPR-related, since their software needed to comply with many other regulations across the world.

P2 was the least experienced out of all the participants, with about 5 years of open-source

experience. They maintained a positive outlook of the GDPR, though they mentioned that they disagreed with how the law handles cookie policy. They believe that most of what is contained in the GDPR is just "common courtesy" that businesses should provide to their consumers.

P3 was the most experienced of all the participants, and had numerous notable experiences to share regarding the evolution of the open-source community over the years. In general, P3 believes that the open-source community's perception of the GDPR is positive due to the fact that much of the GDPR's content aligns with the principles of the community. According to P3, most open-source contributors contribute anonymously, and as such, they highly value the protection and safeguarding of personal identifiable information (PII). Most of what the GDPR mandates were features that P3 planned to implement anyway, with the law simply being a strong motivating factor to actually get it done. However, P3 believes that the open-source landscape is much different compared to industry. Having experience with both, P3 sees that companies, specifically marketing services companies, undergo much more scrutiny by the EU judicial system than the open-source community, and in P3's opinion, for good reason. P3 mentioned experience with testifying in court regarding a potential client they had turned down due to the fact that they had been asked to "cover for the fact that they've been systematically violating HIPAA for three years".

When asked about a GDPR concept they found most difficult to implement, both P2 and P3 asserted that they did not believe compliance to be very difficult. They both broke down their compliance strategy into a few simple steps: P2 provides to the user "easy ways to request, delete, and modify user data" along with an email contact to handle more complex requests, and P3 tries to collect as little PII as possible, and when they do collect data, they tell users exactly what is being collected and how they will use it, as well as providing a way

to remove it. P2 specifically cited the "right to be forgotten" as a major GDPR concept they value. P3 stated that the most difficult part of compliance arises when attempting to incorporate external plug-ins into their project, where the developer is unsure whether or not the plug-in is compliant. Specifically, P3 pointed out the case of Google Analytics, in which most developers believed the tool to be GDPR-compliant until an EU court deemed it not so [8].

Neither P2 nor P3 ever consulted a legal expert for aid in assessing the GDPR compliance in their projects. P2 stated the fact that their projects are non-profit and small in scale, as well as the lack of perceived risk of consequences for non-compliance as reasons for not involving legal aid. P3 also believed that the risk of legal action being taken against their non-profit projects was too low to necessitate aid from a lawyer. P3 also mentioned a lack of budget being a limiting factor in acquiring legal help.

Chapter 6

Discussion

6.1 Interviews

Our interview participants gave us numerous examples to help us answer *Research Question 1*. They mentioned struggles with cookies, embedding content, and integrating external tools with unknown compliance into their projects. The link between these issues is that these are common features implemented in almost every website. As a result, these features are also some of the more well-defined in terms of requirements established by the GDPR. This can actually turn into a pain point for developers, since if they disagree with how the GDPR handles one of these issues, it can become exceptionally frustrating being forced to implement it in that way over and over again with every project. This is a problem that arises when legislators do not consult software developers when creating requirements. Developers can warn lawmakers when they detect a requirement that would lead to unnecessary developmental overhead or would force developers into encoding bad user-experience designs into their project.

To answer *Research Question 3*, P1 seemed to have an extremely different experience and outlook on the GDPR compared to P2 and P3. It seems that the increased liability that comes from implementing GDPR in industry creates frustration and uncertainty when developing compliant software. P2 and P3 both explained how they loosely follow the "spirit"

of the law and do not feel at risk of running into any legal trouble for incorrect compliance.

A major difference between P1 compared to P2 and P3 was their outlook on businesses. In general, P1 holds a generally positive and forgiving viewpoint of them. They believe that most companies have a vested interest in providing secure software: consumer interest is trending towards that area, and companies want to comply with the law. P1 was sympathetic towards Facebook during the Cambridge Analytica scandal, believing Facebook to have never intended for personal data on their platform to be abused in such a way. In contrast, P2 and P3 seem to have a negative and cynical outlook on businesses and their approach to regulation. P2 states that the GDPR's guidelines surrounding cookie policy are often implemented with "dark patterns" by businesses, being intentionally designed to be difficult to opt out of while being much easier to opt in, leading to users being coerced into accepting cookies. P3 believes that businesses will implement as little compliance as they think they can get away with, with the sale of "secure" software essentially boiling down to branding instead of any real novel functionality. This difference in perspective could motivate their differing perception of the GDPR. P1 believes there is enough external incentive for businesses to provide secure software without the GDPR coming into play, while P2 and P3 believe that regulations like the GDPR are fundamentally necessary to prevent personal data from being abused. Overall, open-source developers seem to value privacy strongly and therefore support the core principles expressed by the GDPR, while industry developers are limited by the strict fines and the frustrations of deciphering ambiguous requirements.

Calling back to *Research Question 2*, neither P2 nor P3, our open-source developers, received legal aid for any of their GDPR implementation, while P1 worked with lawyers closely for their industry work. Thus, the stakes for P1 seem to be much higher compared to P2 and P3. P1's interactions with lawyers seemed to be a source of stress and anxiety for them,

as they received a continuous stream of new mandates from their legal team in order to accommodate the emergence of new regulations, as well as altered interpretations of existing laws. Both P2 and P3 mentioned that their open-source projects are non-profit, which seems to influence their impression that they will not encounter any significant penalties for accidental non-compliance. Meanwhile, P1 expressed frustration with the significant penalties incurred by the law for non-compliance, which combined with the loose nature of the regulation, leads to uncertainty and additional work required to parse through EU judicial rulings in order to ensure strict compliance. As such, it would appear that the penalties associated with non-compliance are a clear driving factor in shifting developer perception from positive to negative. Developers have less of an issue with the loose nature of the law when they are not in constant stress and fear of legal repercussion.

6.2 Quantitative Analysis

To answer *Research Question 4*, our analysis showed that GDPR-related pull requests had significantly more comments, commits, and additions, and were active for longer compared to their non-GDPR counterparts on the same repository, while non-GDPR PRs had significantly more changed files than GDPR PRs. More comments suggests that more discussion between GitHub collaborators was required for GDPR implementation, and a longer active time also suggests a longer period of work and deliberation required to complete GDPR PRs. More commits suggests more cycles or iterations of changes, which could mean that developers were either implementing lots of different features in the same PR, or they had to do multiple iterations of the same change in order to refine implementation. We can also

speculate about why we see significantly more additions but not deletions. Implementing GDPR compliance usually does not involve removing any functionality - it involves adding additional features to give more control to the user over their personal data. As for the increased changed files in non-GDPR PRs, this is more difficult to explain, as we know very little about the actual content within the non-GDPR PRs. A possible explanation is that when implementing GDPR features, developers choose to group all of the features together into just a few files for organizational purposes.

When comparing the merged, closed, and open status of the pull requests, we found that a large majority of PRs in both datasets were open. We do not know why the sample we collected had almost no merged PRs, for we were able to find ample examples of merged PRs from our preliminary searching. It is possible that the GitHub search algorithm prioritizes open PRs when returning results, since developers who use the GitHub search engine would most likely be more interested in open PRs that they can actually comment on and contribute to, unlike closed PRs that cannot be interacted with in any fashion. However, without more information about the inner workings of GitHub's search algorithm, we can only speculate.

6.3 Interviews vs. Quantitative results

Interestingly, the testimonies of P2 and P3 seem to contradict the quantitative metrics we gathered from our GitHub searching. Although GDPR-related PRs seem to involve more activity than other PRs, our two open-source developers did not perceive any significant impact to their software development process, and overall had a very positive viewpoint of the GDPR. Both developers were not particularly concerned with the finer details of the law and

did not see any significant legal risk to themselves for potential non-compliance, which may have influenced their positive perception. The negative aspects of the law referenced by P1, such as the ambiguous interpretation, were non-issues to P2 and P3 because they did not attempt strict compliance. One important detail of note is that not all open-source projects are non-profit, so if we conduct more interviews in the future, we may find that more open-source developers may experience similar pains to P1's experience implementing GDPR in industry.

Despite asserting that GDPR did not affect them in any significant way in response to our direct questioning, P2 and P3's responses to some of our other questions may provide illumination to this discrepancy. P3 stated that many of the GDPR features they implemented were privacy features that they had already planned but had neglected in favor of functional requirements. When the GDPR went into effect, this was a catalyst to motivate them to finally act on their plans for privacy implementation. This scenario provides a plethora of explanations for why we see longer active times and more comments, commits, and additions in GDPR-related pull requests. If developers initiate GDPR implementation, but then delay it in order to focus on other issues, this lengthens the "active time" metric we were examining. Plans for general privacy implementation that developers then switch to direct GDPR implementation may require more discussion and code changes as developers switch their design approach.

Therefore, the developers' positive perceptions of the GDPR may have clouded their retrospective judgements on the actual work that they put in to implement the law, just as P1's negative perception may have caused them to overemphasize the work that was required for their implementation. Perceptions seem to play a very significant role in influencing a developer's analysis of their work, even from an objective, analytical standpoint. This reinforces our belief that legislators should consult closely with software developers when writing

legislation that is relevant to them. In this way, legislators can form realistic expectations that developers can feel satisfied about fulfilling, leading to a more positive development experience.

6.4 Limitations

Our interviews were limited by sample size in particular (n=3). With only three participants, it is difficult to paint a comprehensive picture of the open-source community as a whole. This limitation was caused by a lack of time, and our strategy for contacting GitHub users for interviews may have been suboptimal. We received some feedback in response to our emails from two recipients who considered our emails to be spam, rather than legitimate research. An improved method of establishing initial communication with users should be considered.

A limitation with our quantitative analysis of GitHub pull requests is the automated nature in which we collected our data. Due to our large sample size in this area, we know little about the actual content within the pull requests we acquired. In our GDPR-related pull requests, we do not know what kind of compliance they may have been implementing and we cannot ascertain the quality or accuracy of the implementation as it pertains to the law. In addition, our own lack of legal counsel as researchers prevented us from making any real judgements in this regard.

6.5 Future Work

To improve our interviews, a huge step would be simply to get more participants. With a larger sample size, we can be more certain about the conclusions we draw from the responses. To expand upon our work with GitHub searching, we could complete a sentiment analysis of developer comments on GitHub to try to glean some perception from users on the website itself. We could also involve legal experts on the GDPR to aid us in assessing the quality of GDPR implementations on GitHub.

We could also examine data privacy laws other than the GDPR like the California Consumer Privacy Act. We could compare such laws to the GDPR and compare their efficacy through additional metrics. If our goal is to improve future legislation, it would be helpful to learn the successes and failures of US laws that have drawn inspiration from the GDPR.

In our research, we looked at the impact of the GDPR in two major areas: the software and software developers. An untapped area for future work is to survey the users of GDPR-compliant systems. Our interview participants mentioned how GDPR requirements can force developers to implement features that they believe will worsen the user experience, such as obtrusive cookie consent pages. By surveying users about their experiences, we can learn how the GDPR affects individuals from across the spectrum.

Chapter 7

Conclusions

In the United States, user data is viewed as a commodity to be bought and sold. Once users agree to allow their personal information to be collected, companies have free reign to use that data as they see fit. The GDPR appeared as a challenge to this mindset that had become so prevalent in the US. Unparalleled in its impact and scope, developers have spent countless hours trying to adapt their systems to be GDPR-compliant. With stark penalties for failing to comply, we as researchers wondered about the impact this law has had on the open-source landscape. In open-source development, anyone can contribute to a project, and contributors may be anonymous. We wished to examine how open-source developers implement GDPR requirements despite the risk of detrimental legal recourse.

In our research, we conducted a quantitative analysis of GitHub pull requests, comparing GDPR-related pull requests to non-GDPR pull requests from the same repositories. We also conducted a small pilot interview study with three participants with backgrounds in both open-source and industry. In our quantitative study, we found that GDPR-related pull requests involved significantly more activity in a majority of the parameters we examined, those being comments, active time, commits, and additions. In our interviews, our participants with open-source experience held a much more positive perception of the GDPR compared to our participant with industry experience. Our open-source participants had a much lower perceived risk associated with implementing GDPR compliance, and followed

loose rules to guide their GDPR implementation, never consulting legal guidance to aid this process. Meanwhile, our industry participant explained working closely with a legal team constantly to ensure strict compliance, and experienced much more frustration when it came to complying with the law. In this way, we find that the GDPR has a significant impact on open-source development, requiring more activity than other software features to complete, but at the same time, open-source developers do not believe the GDPR has significantly affected their development process. We find that software developers in industry engage with the GDPR in a much more formal manner and as such are much more critical of its content.

Thus, we find the open-source community to be a generally accepting environment of the GDPR. The lack of liability allows the finer principles of data privacy and user prioritization to flourish, while the downsides, such its ambiguous interpretation and heavy fines, are not as impactful. Further questions may be raised as to whether or not this perceived lack of risk is well-founded, or whether the increased activity on GDPR-related pull requests has a negative impact. We remain optimistic about the role that data protection laws can play to give everyday individuals more autonomy over their personal information on the internet, and the positive reception we received from open-source developers is encouraging.

Bibliography

[1] URL https://edps.europa.eu/data-protection/data-protection/legislation/history-general-data-protection-regulation_en.

[2] URL https://www.enforcementtracker.com/.

[3] . URL https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/lawful-basis-for-processing/consent/.

[4] . URL https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/individual-rights/.

[5] URL https://legal.thomsonreuters.com/en/insights/articles/top-five-concerns-gdpr-compliance.

[6] Mar 2018. URL https://gdpr-info.eu/art-83-gdpr/.

[7] May 2019. URL https://edpb.europa.eu/news/news/2019/1-year-gdpr-taking-stock_pl.

[8] Apr 2023. URL https://usercentrics.com/knowledge-hub/google-analytics-and-gdpr-compliance-rulings/.

[9] Yaqoob Al-Slais. Privacy engineering methodologies: A survey. In *2020 International Conference on Innovation and Intelligence for Informatics, Computing and Technologies (3ICT)*, pages 1–6, 2020. doi: 10.1109/3ICT51146.2020.9311949.

[10] Abdulrahman Alhazmi and Nalin Asanka Arachchilage. I'm all ears! listening to software developers on putting gdpr principles into software development practice. *Personal and Ubiquitous Computing*, 25(5):879–892, 2021. doi: 10.1007/s00779-021-01544-1.

[11] Robert Bateman. How the ccpa (cpra) is similar to the gdpr, Feb 2023. URL https://www.termsfeed.com/blog/ccpa-similar-gdpr/.

[12] Kathrin Bednar, Sarah Spiekermann, and Marc Langheinrich. Engineering privacy by design: Are engineers ready to live up to the challenge? *The Information Society*, 35 (3):122–142, 2019. doi: 10.1080/01972243.2019.1583296. URL https://doi.org/10.1080/01972243.2019.1583296.

[13] Tek Raj Chhetri, Anelia Kurteva, Rance J. DeLong, Rainer Hilscher, Kai Korte, and Anna Fensel. Data protection by design tool for automated gdpr compliance verification based on semantically modeled informed consent. *Sensors*, 22(7):2763, 2022. doi: 10.3390/s22072763.

[14] Nicholas Confessore. Cambridge analytica and facebook: The scandal and the fallout so far, Apr 2018. URL https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html.

[15] J.B. Earp, A.I. Anton, L. Aiman-Smith, and W.H. Stufflebeam. Examining internet privacy policies within the context of user privacy values. *IEEE Transactions on Engineering Management*, 52(2):227–237, 2005. doi: 10.1109/TEM.2005.844927.

[16] Sheila Jambekar. Gdpr: Data subjects, controllers and processors, oh my!, Oct 2017. URL https://www.twilio.com/blog/2017/10/gdpr-data-subjects-controllers-processors.html.

[17] Ze Shi Li, Colin Werner, and Neil Ernst. Continuous requirements: An example using

gdpr. In *2019 IEEE 27th International Requirements Engineering Conference Workshops (REW)*, pages 144–149, 2019. doi: 10.1109/REW.2019.00031.

[18] Natasha Lomas. France spanks google $170m, facebook $68m over cookie consent dark patterns, Jan 2022. URL https://techcrunch.com/2022/01/06/cnil-facebook-google-cookie-consent-eprivacy-breaches/.

[19] Abhishek Mahindrakar and Karuna Pande Joshi. Automating gdpr compliance using policy integrated blockchain. In *2020 IEEE 6th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*, pages 86–93, 2020. doi: 10.1109/BigDataSecurity-HPSC-IDS49724.2020.00026.

[20] Gareth Mitchell. How much data is on the internet?, Apr 2020. URL https://www.sciencefocus.com/future-technology/how-much-data-is-on-the-internet/.

[21] Supreeth Shastri, Vinay Banakar, Melissa Wasserman, Arun Kumar, and Vijay Chidambaram. Understanding and benchmarking the impact of gdpr on database systems. *Proceedings of the VLDB Endowment*, 13(7):1064–1077, 2020. doi: 10.14778/3384345.3384354.

[22] David William Silva. Gdpr part 1: Context, motivations, and goals, Mar 2022. URL https://idpro.org/gdpr-part-1-context-motivations-and-goals/.

[23] Sean Sirur, Jason R.C. Nurse, and Helena Webb. Are we there yet? understanding the challenges faced in complying with the general data protection regulation (gdpr). In *Proceedings of the 2nd International Workshop on Multimedia Privacy and Security*, MPS '18, page 88–95, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450359887. doi: 10.1145/3267357.3267368. URL https://doi.org/10.1145/3267357.3267368.

[24] N. van Dijk, A. Tanas, K. Rommetveit, and C. Raab. Right engineering? the redesign of privacy and personal data protection. *International Review of Law, Computers amp; Technology*, 32(2–3):230–256, Apr 2018. doi: 10.1080/13600869.2018.1457002.

[25] Branka Vuleta. 10 unbelievable gdpr statistics in 2023, Mar 2023. URL https://legaljobs.io/blog/gdpr-statistics/.