An Exploratory Mixed-methods Study on General Data Protection Regulation (GDPR) Compliance in Open-Source Software

Lucas Franke
Virginia Tech
United States
lfranke@vt.edu

Huayu Liang
Virginia Tech
United States
huayu98@vt.edu

Sahar Farzanehpour
Virginia Tech
United States
saharfarza@vt.edu

Aaron Brantly
Virginia Tech
United States
abrantly@vt.edu

James C. Davis
Purdue University
United States
davisjam@purdue.edu

Chris Brown
Virginia Tech
United States
dcbrown@vt.edu

Abstract

Background: Governments worldwide are considering *data privacy regulations*. These laws, such as the European Union's General Data Protection Regulation (GDPR), require software developers to meet privacy-related requirements when interacting with users' data. Prior research describes the impact of such laws on software development, but only for commercial software. Although open-source software is commonly integrated into regulated software, and thus must be engineered or adapted for compliance, we do not know how such laws impact open-source software development.

Aims: To understand how data privacy laws affect open-source software (OSS) development, we focus on the European Union's GDPR, as it is the most prominent such law. We investigated how GDPR compliance activities influence OSS developer activity (RQ1), how OSS developers perceive fulfilling GDPR requirements (RQ2), the most challenging GDPR requirements to implement (RQ3), and how OSS developers assess GDPR compliance (RQ4).

Method: We distributed an online survey to explore perceptions of GDPR implementations from open-source developers (N=56). To augment this analysis, we further conducted a repository mining study to analyze development metrics on pull requests (N=31,462) submitted to open-source GitHub repositories.

Results: Our results suggest GDPR policies complicate OSS development and introduce challenges, primarily regarding the management of users' data, implementation costs and time, and assessments of compliance. Moreover, we observed negative perceptions of the GDPR from OSS developers and significant increases in development activity, in particular metrics related to coding and reviewing, on GitHub pull requests related to GDPR compliance.

Conclusions: Our findings provide future research directions and implications for improving data privacy policies, motivating the need for relevant resources and automated tools to support data privacy regulation implementation and compliance efforts in OSS.

ESEM '24, October 24–25, 2024, Barcelona, Spain

ACM ISBN 979-8-4007-1047-6/24/10
https://doi.org/10.1145/3674805.3686692

CCS Concepts

• **Social and professional topics** → **Governmental regulations**;
• **Software and its engineering**;

Keywords

Data Privacy, Regulatory Compliance, Open-Source Software

ACM Reference Format:
Lucas Franke, Huayu Liang, Sahar Farzanehpour, Aaron Brantly, James C. Davis, and Chris Brown. 2024. An Exploratory Mixed-methods Study on General Data Protection Regulation (GDPR) Compliance in Open-Source Software. In *Proceedings of the 18th ACM / IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM '24), October 24–25, 2024, Barcelona, Spain.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3674805.3686692

1 Introduction

Software products collect an increasing amount of data from users to enhance user experiences through personalized, machine learning-enabled [52] application behaviors [33] and marketing [78]. Such practices may benefit users, but also threaten their well-being. For example, in 2013, Facebook allowed the political research firm Cambridge Analytica to access data on ~87 million Facebook users [61]. This data was used to influence US elections [115, 116].

To protect their citizens, over 100 governments worldwide are developing *data privacy regulations* [106]. Their goal is to constrain how their citizens' personal data is collected, processed, stored, and saved. Some target specific industries, *e.g.,* the United States's Health Insurance Portability and Accountability Act (HIPAA), which places requirements on healthcare organizations handling medical data [7]. Others cover personal data regardless of context, *e.g.,* the European Union's General Data Protection Regulation (GDPR), which grants rights to EU citizens and affects entities that handle their data [12]. The penalties for non-compliance with data privacy laws and regulations may be severe [18, 46]. For example, under GDPR, corporations have been fined millions or billions of euros [79]. Most organizations store and manipulate this data electronically through software, and so ensuring the software is in legal compliance is an important software engineering task.

Data privacy regulations are a challenging source of software requirements because they entail both technical and legal expertise. Developers must implement required features, such as obtaining consent from users for data collection, to ensure their organizations' products are compliant. However, developers have limited legal

knowledge [80, 110] and receive minimal training [21, 54]. This can lead to coarse solutions, such as exiting the affected market [87]—hundreds of websites simply banned European users when GDPR went into effect [95, 104]. Research explores the impact of data privacy regulations on businesses [71, 72, 87], users [22, 32, 67], and observable software product properties such as website cookies [66] and database performance [91]. While the question of how such laws affect software development has received much attention in recent years, these studies concentrate on the development of commercial products and services [20, 29]; we lack knowledge of the effects of GDPR on open-source software (OSS) development.

The goal of this work is to describe the impact of data privacy regulation compliance on open-source software. Our study is the first on this topic.[1] We therefore adopt an exploratory methodology to provide an initial characterization and identify phenomena of interest for further study. Our study draws on two data sources collected in two phases. The first phase consists of a *sample survey* [98] gathering insights from developers' experiences with GDPR implementations in OSS, collected via an online survey (N=56). To further investigate the impact of GDPR in OSS, the second phase involves *mining software repositories* [98] to collect and analyze activities in open-source projects on GitHub. We examine metrics and sentiments on 31,462 pull requests (PRs), divided into 15,731 GDPR and non-GDPR PRs.

Our results show GDPR compliance negatively impacts open-source development—incurring complaints from developers and significantly increasing coding and reviewing activities on PRs. In addition, despite the benefits of data privacy regulations for users, we find developers have mostly negative perceptions of the GDPR, reporting challenges with implementing and verifying policy compliance. We also find that interactions with legal experts hinder development processes, yet developers rarely consult with legal teams—often relying on ad hoc methods to verify GDPR compliance.

In sum, our contributions are:

- We survey OSS developers to understand developers' experiences with GDPR compliance and challenges with implementing and assessing data privacy regulations.
- We empirically analyze the impact of GDPR-related implementations on development activity metrics.
- We use natural language processing (NLP) techniques to evaluate the perceptions of GDPR compliance through discussions on OSS repositories.

Significance: This work contributes an exploratory analysis on the impact of GDPR compliance on open-source software. It identifies interesting phenomena for further research—in particular opportunities to support policy implementation and verification. We also provide recommendations for policymakers and software developers to improve data privacy regulations and their implementation.

2 Background

2.1 Software Regulatory Compliance

2.1.1 In General. Software requirements are divided into two categories: functional and non-functional [94]. Functional requirements pertain to input/output characteristics, *i.e.,* the functions the software computes. Non-functional requirements cover everything else, such as resource constraints, deployment conditions, and development process. One major class of non-functional requirement is *compliance with applicable standards and regulations.* These requirements are typically developed and enforced on a per-industry basis in acknowledgment of that industry's risks and best practices [53].

Complying with standards and regulations has been part of software engineering work for many years. Some standards apply to any manufacturing process, *e.g.,* the ISO 9001 quality standard [11]. Others are generic to software development (*e.g.,* ISO/IEC/IEEE 90003 [10]). Still others are contextualized to the risk profile of the usage context, *e.g.,* ISO 26262 [13] or IEC 61508 [9] which describe standards for safety-critical systems [53]; the US HIPAA law (Health Insurance Portability and Accountability Act) which describes privacy standards for handling medical data [7]; and the US FERPA law (Family Education Rights and Privacy Act) which describes privacy standards for handling educational data [5]. Although these regulations are not new (*e.g.,* FERPA dates to 1974, HIPAA to 1996, and IEC 61508 to 1998), software engineering teams still struggle to comply with them [34, 40, 43, 74].

2.1.2 In Open-Source Software. This study focuses on GDPR compliance in open-source software. The reader may be surprised that regulatory compliance is a factor in OSS development, as open-source licenses such as MIT [3], Apache [8], and GNU GPL [6] disclaim legal responsibility. For example, the MIT license, the most common on GitHub [27], states *"the software is provided 'as is', without warranty...[authors are not] liable for any claim, damages, or other liability"*. However, users and developers of OSS may desire regulatory compliance. We note three examples. (1) A majority of OSS is developed for commercial use [47] and may require standards or regulatory compliance [109]. (2) Users with OSS components in software supply chains [51, 82] may request compliance requirements such as web cookies. The developers may service these requests. (3) Users may extend open-source software themselves and undertake their own compliance analysis [97]. Standards such as IEC 61508–Part 3 include provisions for doing so [59].

Open-source software is no longer a minor player in commercial software engineering. Multiple estimates suggest that OSS components comprise the *majority* of many software applications [47, 81]. In a 2023 survey of ~1700 codebases across 17 industries, Synopsys found OSS in 96% of the codebases and reported an average contribution of 75% of the code in the codebase [100]. It is therefore important to understand how OSS development considers non-functional requirements such as regulatory compliance.

2.2 Privacy Regulations, Especially GDPR

2.2.1 Consumer Privacy Laws. In §2.1 we discussed standards and regulatory requirements that affect software products based on industry. Recently a new kind of regulation has begun to affect software: consumer privacy laws. The most prominent example of such a law is the European Union's General Data Protection Regulation (EU GDPR), enacted in 2016 and enforceable beginning in 2018. Examples in the United States include the California Consumer Privacy Act (CCPA, enacted 2018) and the Virginia Consumer Data Protection Act (CDPA, enacted 2021). Similar legislation has been considered by >100 governments [58, 106].

[1] This paper is an extension on our preliminary work, presented as a poster [44].

2.2.2 The General Data Protection Regulation (GDPR). The General Data Protection Regulation (GDPR) [12] protects the personal data of European Union (EU) citizens, regardless of whether data collection and processing is based in the EU. The law has implications for entities that interact with the personal data of EU citizens, divided into data *subjects*, data *controllers*, and data *processors* [45]. Data subjects are individuals whose personal data is collected. Data controllers are any entities —organization, company, individual, or otherwise — that own, control, or are responsible for personal data. Data processors are entities that process data for data controllers. The GDPR grants data subjects rights to their personal data, providing guidelines and requirements to data controllers and processors to understand how to properly handle this data.

GDPR compliance is complex for software engineers and consequential for their organizations. Data controllers and processors commonly use software, *e.g.,* a controller's mobile app transmits data to its backend service and processors subsequently access and update the database. Software teams must determine appropriate data policies, update their systems to comply, and validate them, *e.g.,* incorporating cookie consent notices into websites to provide users with informed consent [107]. Anticipating a lengthy compliance process, the EU enacted the GDPR in 2016 but made it enforceable in 2018, allowing two years for corporations to prepare [1]. Companies in the US and UK alone invested $9 billion in GDPR compliance [111]. As of December 2022, many use manual compliance methods or are not compliant [14]. Non-compliance is costly: thousands of distinct fines have been imposed on non-compliant data controllers and processors, exceeding €2.5 billion [15].

Although GDPR compliance affects any software that processes the data of EU citizens, and open-source software components comprise the majority of many software applications that process such data [47, 81, 100], *to the best of our knowledge there is no prior research on the impacts of GDPR compliance in open-source software.*

3 Methodology

3.1 Data Availability and Research Questions

In §2 we described a range of privacy-related standards and regulations. We noted that there has been little study of the effect of these requirements on OSS development. To address this gap, we need data. Table 1 estimates the availability of data associated with these requirements through two common metrics: the number of posts on Stack Overflow and the number of pull requests on GitHub.

Table 1: Software engineering data availability for privacy legislation. Data from keyword search on Nov. 13, 2023. We studied GDPR.

Privacy Law (Year)	Stack Overflow	GitHub-PRs
GDPR (2016)	2058	64 K
HIPAA (1996)	725	5 K
CCPA (2018)	96	1 K
FERPA (1974)	35	254
CDPA (2021)	7	19
PIPEDA (2000)	5	31

Based on this data, we scoped our study to the EU's GDPR; and to open-source software hosted on GitHub, currently the most popular hosting platform for OSS. We answer four research questions:

RQ1: How does GDPR compliance influence development activity on OSS projects?
RQ2: How do OSS developers perceive fulfilling GDPR requirements?
RQ3: What GDPR concepts do OSS developers find most challenging to implement?
RQ4: How do OSS developers assess GDPR compliance?

We analyzed data from quantitative and qualitative sources: surveying open-source developers and mining OSS repositories on GitHub. We integrate this data in answering RQ1 and RQ2, and use the survey data alone to answer RQ3 and RQ4.

3.2 Data Source 1: Developer Survey

To explore the impact of implementing GDPR policies on OSS development, we distributed an online survey for open-source developers. This data informed our answers to all RQs. We used a four-step approach motivated by the framework analysis methodology [89] for policy research to collect and analyze data in the second phase of our experiment. An overview of this process is presented in Table 2. Our **Institutional Review Board (IRB)** provided oversight.

3.2.1 Step 1: Pilot Study and Data Familiarization. To formulate an initial thematic framework for our qualitative analysis, we conducted semi-structured pilot interviews with OSS developers ($n = 3$). As no prior work has explored the perceptions of GDPR compliance in OSS, pilot interviews gave us insight into developers' perceptions and experiences with implementing GDPR concepts in the context of open-source software development. Two subjects had contributed to PRs in our dataset, and the third was a personal contact. They had a wide range of open-source development experience, from < 1 year to > 20 years. Interviews were transcribed using Otter.ai and coded by two researchers to inform our survey.

Thematic analysis of our pilot interviews highlighted challenges with implementing GDPR requirements in open-source software. One participant worked at a large corporation and outlined differences between GDPR compliance at their company and in OSS, namely with (1) approaches used to assess whether compliance is implemented correctly, and (2) access to legal teams. The other two participants discussed the impact of the GDPR, noting its privacy benefits as well as challenges with implementing requirements and assessing compliance. These findings informed our survey.

3.2.2 Step 2: Survey Design. The survey consisted of open-ended and short answer questions seeking details about GDPR implementation and experiences in the context of open-source software development. We used the pilot study interview results to identify topics to focus on in the survey. Based on the interviews, we asked about the perceived impact of the GDPR on data privacy, the most difficult concepts to implement, and how they assess GDPR compliance. The survey instrument is in the supplemental material.

3.2.3 Step 3: Participant Recruitment. We distributed our survey in three rounds. In the first round, we emailed a sample of 98 developers who authored or commented on GDPR-related PRs with publicly available email addresses. We received 5 responses, *i.e.,* a 5% response rate. In the second round, we made broader calls for participation on Twitter and Reddit. We received 44 responses, 2 of which indicated no experience implementing GDPR compliance.

Table 2: Overview of sample questions from pilot interview study and survey design/analysis for framework analysis approach used for Data Source 2. The final column notes the inter-rater agreement score for these themes using the κ score, prior to reaching agreement.

Interview Question	Codes	Survey Question	Codes	κ
What meaningful impact, if any, do you believe the GDPR has had on data security and privacy?	data privacy, rights to users, data collection	What impact, if any, do you believe the GDPR and similar data privacy regulations have had on data security and privacy?	data privacy, data processing, data collection, insufficient information, data breach, fines	0.736
What GDPR concepts do you find the most difficult or frustrating to implement?	None, data minimization, embedded content	What GDPR concepts do you find the most difficult or frustrating to implement?	privacy by design, data minimization, cost, data processing, user experience, data management, security risks, None, lawfulness and dispute resolution, time, right to erasure	0.929
Have you had to specifically seek out legal consultation on GDPR-related issues, and if so, how did that affect your development process?	Yes/No; no effect, negative effect (time)	Have you had to specifically seek out legal consultation on GDPR-related issues, and if so, how did that affect your development process?	Yes/No; N/A, no effect, positive effect, negative effect (cost, time, data storage, data processing,...)	0.514
During your software development projects, do you frequently consult with a legal team, and if so, how does this impact the development processes? If not, how did you assess GDPR compliance for your software projects?	Yes: legal consultation; No: privacy by design, data minimization	During your software development projects, have you consulted with a legal team? If not, how do you assess GDPR compliance for your software projects?	Yes: legal consultation; No: accountability system, online resources, self-assessment, data management, none), N/A	0.668
—	—	Has implementing GDPR concepts for compliance impacted your development process in any way? (*yes/no/maybe*) Please explain:	positive impact (logging, privacy by design), negative impact (cost, data management, security,...), no impact	0.860

All survey respondents in these rounds were entered in a drawing for two $100 Amazon gift cards. After a few months, we undertook the third round, redistributing our survey to an additional 235 GitHub users with GDPR implementation experience (authored GDPR-related PRs in our dataset) and offered individual compensation ($10 gift card) to encourage participation. We received 9 responses (4% response rate). In total we have data from 56 survey participants (14 from GitHub and 42 from Twitter/Reddit).

Our participants have a median of approximately 5 years of OSS development experience (avg = 5.9) and 6 years of general industry experience (avg = 7.7). Participants reported contributing to a variety of OSS projects such as Mozilla, Wordpress, Fedora, Moodle, Ansible, Flask, Django, Kubernetes, PostGreSQL, OpenCV, GitLab, and Microsoft Cognitive Toolkit.

3.2.4 Step 4: Data Analysis. To analyze our survey results, we used an open coding approach. Two researchers independently performed a manual inspection of responses–highlighting keywords and categorizing responses based on the pre-defined themes derived from our pilot study. If new themes arose, the coders discussed and agreed upon adding the new theme. Then, both coders came together to merge their individual results. Finally, we used Cohen's kappa (κ) to calculate inter-rater agreement (see Table 2).

3.3 Data Source 2: GDPR PRs on GitHub

We collected data concerning GDPR compliance by analyzing pull requests on GitHub repositories. Pull requests are a mechanism on GitHub that allow developers to collaborate on open-source repositories, involving code contributions from developers to be reviewed and merged into the source code.[2]

3.3.1 GDPR and non-GDPR PRs. We used the GitHub REST API to search for *GDPR-related pull requests*—pull requests returned by the GitHub API's default search with the query string "GDPR". Manual inspection suggested the results are typically English-language PRs related to (GDPR) data privacy regulatory compliance.

Using this method, we collected GDPR-related PRs created from April 2016 (when the GDPR was adopted by the European Parliament) to January 2024. We removed content submitted by users with "bot" in their username [16] and designated as a bot type according to the GitHub API[3] to avoid PRs generated by automated systems. This resulted in 15,731 GDPR-related pull requests across 6,513 unique GitHub repositories. For comparison, we also collected a random sample of 15,731 pull requests created in these same repositories after April 2016 that did **not** mention "GDPR", which we call *non-GDPR-related pull requests*. The studied repositories had a median of 14 stars (avg = 1,635), 11 forks (avg = 416), 727 commits (avg = 8,997), 172 PRs (avg = 1,425), and 15 contributors (avg = 59), suggesting popular, active repositories. The distribution of PRs across all repositories in our GDPR-related and non-GDPR-related datasets is summarized in Table 3.

3.3.2 Measuring Development Activity. To analyze GDPR's impacts, we collected development activity metrics [48] per pull request:

- *Comments:* the total number of comments

[2]https://help.github.com/en/articles/creating-a-pull-request
[3]https://docs.github.com/en/graphql/reference/objects#bot

Table 3: Distribution of PRs in Datasets.

Dataset	min	50%ile	75%ile	90%ile	max
GDPR	1	1	2	3	956
non-GDPR	1	2	10	34	203

- *Active time:* the amount of time the PR remained active (until merged or closed)
- *Commits:* the total number of commits
- *Additions:* the number of lines of code added
- *Deletions:* the number of lines of code removed
- *Changed files:* the total number of modified files
- *Status:* outcome of PR (merged, closed, or open)

We selected these metrics to analyze development activity, specifically to derive coding and code review tasks from pull requests. We compared the distributions of these metrics between GDPR-related and non-GDPR-related PRs using a Mann-Whitney U test, to compare nonparametric ordinal data between the datasets [75]. To control for multiple comparisons on the same dataset, we calculate adjusted p-values using Benjamini-Hochberg correction [30]. We measure effect size (r) for significant results using Cohen's d [39].

3.3.3 Measuring Developer Perception. To augment our survey results, we applied sentiment analysis—a technique to automatically infer sentiment from natural language—on the title, body, commit messages, review comments, and discussion comments from pull requests in our datasets to examine developer perceptions of GDPR compliance. Prior studies have similarly inferred developer sentiment and emotion from GitHub activity, including PR discussion comments [86], review comments [56], commit messages [49], and bodies [83]. While this technique sometimes has negative results in software engineering contexts [63], we use it in our exploratory work as a proxy to obtain preliminary insights into developers' sentiments regarding GDPR compliance in OSS.

We followed standard NLP preprocessing steps [68]: (1) We removed bot-generated content using the process described in Section 3.3.1. (2) We removed non-sentiment material: hyperlinks and mentions ("@username"). (3) We tokenized text using the Natural Language Toolkit (NLTK) `tokenize` library. (4) We converted tokens to lowercase and removed punctuation. (5) We removed stopwords such as "but" and "or" (`nltk.corpus` library). (6) We lemmatized the text, *i.e.,* reducing words to their base form (*e.g.,* "mice" becomes "mouse" [23]) using `WordNetLemmatizer` from the `nltk.stem` library. (7) We normalize the data by removing meaningless tokens, such as SHA or hash values for commits, and non-standard English words, such as words that contain numerical values (*i.e.,* "3d") [96].

After preprocessing the data, we were left with 15,731 titles, 14,515 bodies, 15,217 commit messages, 4,922 review comments, and 4,862 discussion comments across the GDPR-related pull requests. We compared these against non-GDPR-related PRs, for which we had 15,731 titles, 13,718 bodies, 15,652 commit messages, 3,427 review comments, and 3,165 discussion comments.

To perform sentiment analysis, we use three state-of-the-art models: Liu-Hu [55], VADER [57], and SentiArt [62]. We fed the preprocessed textual data to each model, which provided compound sentiment scores. We use a t-test (t) to statistically analyze sentiment across our datasets. Moreover, we aim to assess the impact of the GDPR on developer sentiment over time. To accomplish this, we divided the GDPR and non-GDPR PRs into 3-month segments based on the creation date of the PR. Then, we performed sentiment analysis on the binned data to observe whether and how developer sentiments manifest in OSS interactions over the lifecycle of the GDPR regulation — from its initial adoption in 2016, enforcement in 2018, and to the present. We combined all preprocessed textual elements (title, body, commit messages, review comments, and discussion comments) to observe the overall trends in PR communications and compare with non-GDPR data as a baseline sentiment in developer communications for the projects studied.

4 Results

We are interested in understanding the impact of GDPR implementations on open-source software by analyzing development activity and developer perceptions, including challenges with implementation and assessment of compliance. In this work, we answer our research questions using multiple sources—analyzing GitHub repositories and surveying open-source developers. For RQ1 and RQ2, we report views from the survey and the GitHub measurements. For RQ3 and RQ4, we use data only from the survey.

4.1 RQ1: Development Activity

4.1.1 Survey. We surveyed 56 OSS developers to understand the impact of GDPR implementations on development activity. Most participants (n = 41, 73%) responded "Yes" to a question regarding the impact of implementing GDPR concepts on development processes, indicating data privacy compliance effects open-source development. When asked to elaborate, 23 developers provided examples of development impacts related to the GDPR.

Data Management: 11 participants mentioned GDPR requirements increase development efforts. For instance, responses indicated handling personal data (P17) and anonymization (P19), managing data controllers (P21) and data recipients (P23), implementing functionality to limit the collection of personal data (P26), and the monitoring of data subjects from the EU (P28) impact development processes. P53 added "*we had to separate in a clear way sensitive data from the other data*", exemplifying the effort needed to implement compliant data processing in OSS.

Time and Costs: Five participants mentioned GDPR compliance increases development time and costs in OSS. For example, regarding time, respondents said "*it does slow down our development cycle*" (P54) and "*we lost a complete year to be ready*" (P56). For costs, participants said "*budgets have soared*" (P5) and "*costs of production should not go over the cost of consequence of data breach*" (P46).

Design: Three participants also noted the effects of GDPR compliance on the design and structure of software products. For example, P54 responded "*we have to check whether we comply with GDPR every time we draft a new design*" and P55 added "*the design of systems now incorporates the concept of needing to remove PII after the fact*". P21 explained how GDPR compliance reduced the quality of their application's design–replying "*the principle of minimum scope was not observed*"—indicating potential unnecessarily extended scopes of variables in the code [36].

Organization: Three participant responses embodied the negative effects of data privacy regulations on their organization, stating

the GDPR has a "*major impact*" requiring "*an overhaul of project management and program priorities*" (P1). P45 highlighted that "*making sure to follow privacy by design*" is challenging for GDPR compliance in OSS development. One participant also mentioned additional steps to verify implementations affected their development, stating "*we need to make an additional review with the GDPR consultants that functionality that is related to the users' data*" (P53).

Benefits: One participant mentioned benefits to their development team and processes regarding the implementation of GDPR concepts, stating it helped highlight "*things we had not considered before*", such as ensuring that "*logging functionality*" and "*access restrictions*" were in place (P1). However, the majority of responses indicate that GDPR compliance often increases development efforts and incurs negative impacts for open-source developers.

4.1.2 Pull Request Metrics. To further observe the impact of GDPR compliance on OSS, we compared metrics for GDPR and non-GDPR related PRs. Table 4 presents these results. Using a Mann-Whitney U test, we found statistically significant differences between GDPR and non-GDPR PRs in the number of comments, active time, number of commits, lines of code added, lines of code deleted, and number of modified files. We also calculate the effect size for these results.

This indicates that incorporating changes related to the GDPR has a major impact on development work, leading to: increased discussions between developers, longer review times, more code commits, and higher code churn. While we observed significant differences exist in pull request metrics between GDPR and non-GDPR PRs, the calculated effect sizes are "small" [70], indicating low practical differences between the groups. Yet, these findings support our survey results from open-source developers purporting that GDPR compliance efforts affect OSS development.

Finding 1: Developers report implementing GDPR compliance negatively affects development processes–citing cost, time, and data management as concerns.
Finding 2: PRs related to GDPR compliance have significantly more development activity for coding (*commits, additions, deletions, files changed*) and review (*comments, active time*)

Table 4: GDPR (G) vs. Non-GDPR (non-G) GitHub Activity Metrics.

Characteristic	Type	Median	*p*-value
*Comments**	G	1	$p < 0.0001$
	non-G	1	($U = 1.4E8$, $r = 0.09$)
*Active time (days)**	G	418.05	$p < 0.0001$
	non-G	1.78	($U = 1.4E8$, $r = 0.14$)
*Commits**	G	2	$p < 0.0001$
	non-G	1	($U = 1.4E8$, $r = 0.04$)
*Additions**	G	57	$p < 0.0001$
	non-G	19	($U = 1.5E8$, $r = 0.05$)
*Deletions**	G	7	$p < 0.0001$
	non-G	4	($U = 1.3E8$, $r = 0.05$)
*Changed files**	G	4	$p < 0.0001$
	non-G	2	($U = 1.4E8, r = 0.03$)

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

4.2 RQ2: GDPR Perceptions

4.2.1 Survey. We asked participants their perceptions on the impact of GDPR regulations on privacy. Of participants who responded to this question ($n = 25$), most had negative opinions of the GDPR. Three participants were neutral (*e.g.,* "*N/A*" (P4)). We summarize positive and negative perceptions next.

Negative Perceptions: Despite the utility of data privacy regulations, 22 participants reported negative perceptions of the GDPR. These responses primarily focused on three issues: cost, organizations, and enforcement. For costs, respondents noted that implementing GDPR requirements is expensive. Participants said that compliance is "*costly for many companies*" (P16) is "*too expensive*" (P24), and "*the cost of protection should not go over the cost of consequence of data breach...GDPR [isn't] worth the time*" (P46). P55 also highlighted that "*in general there have been major costs to companies of all sizes*". For organizations, participants reported a negative impact of GDPR on companies and organization stating: "*weakens small and medium-sized enterprises*" (P15); "*threatens innovation*" (P18); "*fails to meaningfully integrate the role of privacy-enhancing innovation and consumer education in data protection*" (P23); and "*in order to be safer than risky useful functionality is removed*" (P52). P46 added that the GDPR is "*a lot of headache...jobs for lawyers at the expense of people who are trying to solve real problems*". For enforcement, one subject said "*there is a large gap in GDPR enforcement among member states* (P17) and another observed "*the trend...is an increase in the number of times and the amount of fines*" (P18). Similarly, P49 described GDPR as "*a big hammer*", but was unsure "*if it has necessarily increased security and privacy at this point*".

Positive Perceptions: Eight participants had positive perceptions of the GDPR, generally stating that GDPR enhances data privacy for users. For example, participants said that "*the risk of incurring and paying out hefty fines has made companies take privacy and security more proactively*" (P30), that GDPR brings "*awareness to the importance about privacy*" (P45), that "*data integrity is ensured*" (P47), and "*customers can now delete their data quite easily*" (P54). Participants also appreciated the increased accountability for corporations in safeguarding users' data—for example one participant stated "*Before GDPR data protection was usually considered only as an afterthought if not an outright joke. Nowadays companies will at least consider what they are doing wrong before violating data protection laws, rather than doing it by accident because no-one even thought about it*" (P50). These responses reflect the intentions of the GDPR — to safeguard the rights of users and their data online.

4.2.2 Sentiment Analysis. We investigated the sentiment of developers implementing GDPR concepts by analyzing PR titles, commit messages, review comments, discussion comments, and bodies. Our overall results are in Table 5. We anticipated a higher percentage of negative comments for GDPR-related pull requests. However, we did not find evidence that GDPR-related PRs have less favorable sentiments from developers. In fact, we found they often had *more* positive sentiments than non-GDPR-related PRs—with two of the three models (Liu-Hu and VADER) indicating a statistically significant difference between the GDPR and non-GDPR sentiment. We speculate two explanations. First, non-GDPR-related PRs represent a broad range of code contributions, which could address a number of issues. Second, we are limited by the capabilities of the sentiment

Table 5: GDPR (G) vs Non-GDPR (non-G) Sentiment Analysis

Test	Type	Mean	Variance	*p-value*
*Liu-Hu**	G	0.43	0.27	$p < 0.0001$
	non-G	-0.04	0.28	($t = -4.05$, $r = 0.22$)
*VADER**	G	0.44	0.04	$p < 0.0001$
	non-G	0.21	0.01	($t = -6.47$, $r = 0.02$)
SentiArt	G	0.39	0.01	$p = 0.1399$
	non-G	0.36	0.002	($t = -1.10$, $r = 0.01$)

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

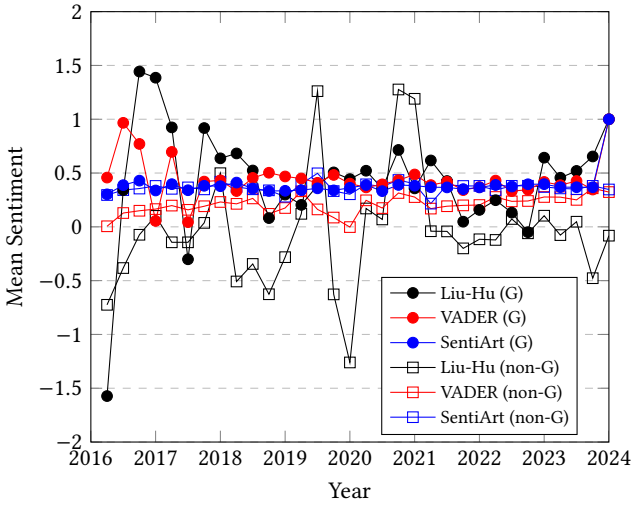

Figure 1: Longitudinal GDPR (G) and Non-GDPR (non-G) Sentiment Analysis Data. We grouped GDPR and non-GDPR data into 3-month segments and used 3 sentiment models. For each model, GDPR data is plotted in a color with a filled marker, and non-GDPR data in the same color but with a hollow marker. The general trend is that sentiment for GDPR data is moderately positive, and more positive than for non-GDPR data.

analyzer. For example, the two most negative commit messages for non-GDPR pull requests said "*obsolete*" and "*fatal*", which are common terms of art in software maintenance tasks [88, 114] (*e.g.,* "*fix fatal error*"). We also observed some variation at the beginning and end of our dataset collection period, but no significant variation in sentiment over time (see Figure 1).

Nonetheless, manual inspection of negatively scored content showed frustrations with GDPR compliance. For instance, one title and commit message described GDPR-related changes to "*avoid lawsuits by mentioning cookies thing*" [90]. Another title stated adding "*just enough EULA [end user license agreement] not to get banned*" [31]. Similar frustrations were shared in a PR body for "*GDPR stuff*" to "*display the annoying cookies banner*" [105]. Discussion comments, such as "*will this conflict with GDPR?*" [24], also highlight OSS developers' confusion with GDPR requirements.

Finding 3: Despite its nominal advantages, most developers had *negative* perceptions of the GDPR and its implementation.
Finding 4: We found developers did *not* express more negative sentiments about GDPR compliance in PR discussions.
Finding 5: Sentiment related to GDPR compliance appears to be stable over time.

4.3 RQ3: Implementation Challenges

In the survey data, we observed three common challenges: data management, data protection, and vague requirements.

Data Management: 11 developers responded that processing and storing users' data according to GDPR requirements is the most challenging concept to implement. For example, participants mentioned challenges implementing "*data protection*" (P24), handling "*personal data*" (P34), the "*exchange of documents containing personal data*" (P32), the "*improper storage*" (30) of user data, and "*knowing what info can or cannot be accessed or saved*" (P49). In particular, four participants mentioned users' right to erasure—or the obligation for data controllers to delete users' data upon request "without undue delay" [4]—as the most complicated requirement to implement. For example, P53 responded, "*it's not always easy enough to implement data processing in a way, that it's anonymized, and if the user would like their data to be erased, be able to continue processing of the results based on user data in an anonymous way*"—describing the complexity of this requirement for their project.

Data Protection: Five participants mentioned security factors as a challenge for GDPR compliance. For instance, participants were concerned with "*data protection*" and "*other security concerns*" (P24), "*leaks*" (P27), and the fact that other entities have "*the ability to steal data*" (P28). P55 noted challenges with handling and securing data in "*central databases, where that data may be relied on by many loosely connected applications and systems*". These responses highlight the difficulties of implementing mechanisms to safeguard users' data.

Vague Requirements: 10 survey respondents highlighted a lack of clear requirements as the biggest challenge with GDPR compliance in OSS. For example, one participant mentioned that GDPR "*is pretty vague*" with a lack of "*standard format*" (P54). Another described confusion in knowing "*how long can data be retained*" and "*what is Personaly[sic] Identifiable Information*"—adding, the "*lack of clarity in the rgulations[sic] leads to confusion*" (P52). Moreover, P48 highlighted the lack of company understanding of GDPR requirements makes compliance difficult.

Beyond these clear categories, we received a wide range of other responses, including "*lawfulness and dispute resolution*" (P47), the conflict between "*individual privacy and the public's right to know*" (P21), and being in a "*rush to regulate*" (P28). P27 mentioned challenges with user experiences, stating "*users endure invasive pop-ups*". P1 noted challenges evolve during the lifetime of a project, stating "*At the beginning of a project, privacy by design and default. In the middle or the end, data minimization and transparency*". Based on the implementation difficulties, participants described limiting functionality—*e.g.,* "*knowing when interacting with EU citizens*" (P49) and "*more than 1,000 news websites in the European Union have gone dark*" (P15). Meanwhile, P17 mentioned difficulties implementing GDPR requirements for data-intensive domains: "*many of the GDPR's requirements are essentially incompatible with big data, artificial intelligence, blockchain, and machine learning*". These challenges motivate new resources to help developers overcome problems related to GDPR implementation and compliance.

Finding 6: The management and protection of user data and vague requirements are key challenges open-source developers face when implementing GDPR requirements.

4.4 RQ4: Compliance Assessment

We found three kinds of responses related to compliance assessment: consulting with legal counsel, referencing other compliance resources, and self-assessment.

Compliance Through Legal Counsel: In our survey results, 15 OSS developers reported consulting with legal teams for GDPR compliance. We were also interested in exploring the impact of seeking legal counsel for GDPR compliance on OSS development processes. Seven participants with experience seeking legal consultations noted that it did have a positive impact on development activity (P6, P13, P14, P45, P53, P55, P56). Participants noted the benefits of seeking legal experts, stating the importance of "*consulting with lawyers on the team who have a seat at the table*" (P45), it "*clarifies requirements and prevents misinterpretations*" (P55), and allowed GDPR compliance to be "*implemented rather easily*" (P56).

However, most participants ($n = 9$) with experience seeking legal counsel lamented the impact, stating: "*it slows things down as code has to be reviewed and objectives revised*" and "*it impacted our approach to the SDLC*" (P1), "*it's a bit of a headache*" (P24), "*it slowed us down...was mostly a box ticking exercise*" (P51), and "*it interrupted the development but it is required*" (P49). Respondents also bemoaned the cost, stating "*for a global project open source project any legal advice would be extremely expensive*" (P52) and "*open-source projects can't afford even to sustain maintainers, not even speaking about legal team*" (P47). P54 noted legal experts also found difficulties with the vagueness of GDPR requirements, replying the "*legal team struggles to interpret how to comply with GDPR, there are a lot of back-and-forth. We have to change our design many times*".

In sum, legal experts can provide valuable insight into data privacy regulations and compliance, but developers often find these interactions negatively impact development processes.

Compliance Resources: To assess GDPR compliance, three participants mentioned a variety of other resources. One participant described formal training on regulatory compliance, with a "*special training on GDPR within the company*" (P16). Another participant responded that their team uses an "*accountability system*" (P24) to assess compliance. Finally, P15 noted using online resources to help, but highlighted their ineffectiveness, stating, "*many of the articles on the Internet about GDPR are incomplete or even wrong".*

Self-assessment: Other developers mentioned they were largely responsible for evaluating the "*legality*" (P18) and "*integrity and confidentiality*" (P23) of the processing and storage of user data in their system on their own. P24 responded developers have to "*consider whether you really need all the data you collect*" while P38 advised to "*get your consent in order*". P53 noted the impact on development teams, stating GDPR implementations "*took us significant amount of time due to several rounds of architecture review*". P18 added there is "*really no good way*" to evaluate compliance.

Finding 7: Developers often do not consult legal experts to validate GDPR compliance, relying on other resources such as compliance training, accountability systems, online resources, and self-assessed data management.
Finding 8: Participants with experience interacting with legal teams provided mixed perceptions, feeling they provided valuable insight but hindered development processes.

5 Discussion and Future Work

Our results demonstrate GDPR-related code changes impact OSS development, significantly increasing development activity with regards to lines of code added and the number of commits included in PRs–indicating increased effort in code contributions and code review activities (§4.1.2). Further, we found GDPR compliance provides a wide range of challenges for OSS development (§4.3) and developers often assess compliance without the help of legal and policy experts (§4.4). These findings posit that implementing GDPR compliance is a challenging activity for OSS developers.

We recognize many stakeholders are involved in adhering to data privacy legislation. For instance, policymakers play a role in data privacy compliance [113]. Data privacy regulations, such as the GDPR, are beneficial for protecting the rights and data of users. However, we noticed developers complained about providing privacy–holding negative perceptions of the GDPR and its implementation. To that end, we provide guidelines to enhance data privacy regulations and software development processes to reduce the negative effects of policy compliance in OSS software.

5.1 Improving Data Privacy Regulations

5.1.1 Provide Clear Requirements. We found developers struggled to implement GDPR concepts (§4.3), and few reported consulting with legal experts to provide insight on policies and assess compliance (§4.4). Thus, most development teams are forced to evaluate the system themselves. Yet, participants complained that understanding compliance is difficult due to the ambiguity of GDPR concepts: for instance, "*the procedure for obtaining user consent and the information provided are unclear*" (P25). Prior work suggests ambiguous [28] and incomplete [38] requirements are inhibit requirements engineering—leading to increased development costs and higher probability of project failure.

To improve program specifications, researchers have explored a variety of techniques. For instance, Wang *et al.* used NLP to automatically detect ambiguous terminology in software requirements [112]. Similar techniques could be applied to regulations such as the GDPR to notify policymakers of unclear language. Additionally, involving software developers in the policy-making process can improve the clarity of policy requirements. Collaboration between policy makers and practitioners improves has been shown to improve policies in public health [37] and education [60]. Verdon argues a good policy must be "understandable to [its] audience" [110, p. 48], yet we observed developers are confused by GDPR requirements. Incorporating developers into data privacy policy-making can provide input on policy requirements and the impact of implementation and compliance on software development.

5.1.2 Policy Resources. We observed OSS developers face challenges implementing GDPR-related changes (§4.3). One participant mentioned formal training on GDPR compliance (P16), however most participants reported compliance assessment resources, such as legal consultations, are largely ineffective for development teams (§4.4). To that end, OSS developers largely resort to implementing and evaluating compliance on their own efforts with "*insufficient information*" (P26). Prior work also outlines issues with software developers and security policies, noting a lack of understanding from programmers [110].

Based on our findings, we suggest creating novel resources to educate developers on policies and their implementation can benefit OSS developers. For instance, we observed GDPR-related PRs incur significantly more OSS development and code review efforts. Prior work suggests contemporary code reviews have additional information needs beyond code, such as specialized expertise and correct understanding for reviewers [84]. Resources to provide information on data privacy-related concepts, such as guidelines or online forums, can further support compliance in OSS. Developers frequently use popular programming-related Q&A websites, *e.g.,* Stack Overflow, to ask questions and seek information online [85]. These resources are also used for discussions on data privacy policy implementation [101] (also see Table 1). However, developers have no way to verify the correctness of responses, which can also become obsolete over time [117]. Further, research shows most privacy-related questions on Stack Overflow lack references to documentation and external resources [102]. Thus, novel resources are needed to provide developers with up-to-date and accurate information regarding data privacy policy implementation.

5.2 Improving Development Processes

5.2.1 Privacy by Design. Participants reported challenges implementing GDPR compliance (§4.3) and negative effects on development practices (§4.1.1). Moreover, our GitHub analysis found GDPR-related changes necessitated significantly more time and effort for developers (see Table 4). Yet, compliance is required for organizations to avoid "*paying out hefty fines*" (P30). Researchers have introduced techniques to better incorporate privacy into development processes—such as Privacy By Design (PBD) [35]. P50 mentioned cultivating "*a privacy-respecting mindset long before GDPR came about*" avoided negative impacts on development processes and made the effort required "quite minimal". However, one survey participant (P1) noted that PBD and prioritizing privacy in software development processes "*requires an overhaul*". Additionally, while PBD can benefit GDPR compliance efforts, Kurtz *et al.* note a scarcity of research in this area and highlight particular challenges with PBD for GDPR implementations, such as ensuring third party libraries also adhere to privacy principles [69].

Studies suggest PBD is effective for new projects (*i.e.,* startups) [103], yet may be ill-equipped for existing projects complying with new and changing data privacy regulations. For instance, Anthonysamy *et al.* demonstrate privacy requirement solutions might solve present issues, but may differ from regulations and policies in the future [25]. More work is needed to explore processes to prioritize data privacy in mature software projects. One solution could be a gradual approach to policy compliance. Some programming languages (*i.e.,* Typescript) support gradual typing to selectively check for type errors in code [92]. Similarly, research in formal methods explores supporting gradual verification of programs [26]. Novel approaches can be used to gradually introduce privacy compliance and assessment in OSS development for evolving software systems.

5.2.2 Automated Tools. We found GDPR compliance impacts OSS development, significantly increasing coding and reviewing tasks on GitHub PRs (see Table 4). Developers who responded to our survey also indicated the impact of GDPR compliance on their project, noting data privacy regulations always need more software (P4), violate the principle of minimum scope (P21), and there is "*no good way*" to assess compliance (P18). These findings point to an increased burden and effort on OSS developers to implement and review GDPR requirements to comply with data privacy regulations and avoid penalties for non-compliance (*e.g.,* losing market share).

To that end, we posit automated tools can reduce the burden of GDPR implementation efforts. One participant mentioned using a tool, an "*accountability system*" (P24), to help assess compliance–however did not provide any details about this system. Our findings for RQ1 (§4.1) show GDPR-related pull requests have significantly more coding involved, consisting of more commits and lines of code added in code contributions, as well as requiring significantly more comments and time in reviewing processes. Prior work has explored systems to support data privacy implementation. For example, Ferrara and colleagues present static analysis techniques to support GDPR compliance [42]. Further efforts can investigate tools to support review efforts for policy-relevant code to streamline compliance assessments. Future systems could also provide automated feedback to developers and reviewers on data privacy regulation compliance. For instance, using NLP techniques [17] or rule-based machine learning approaches [50] to automatically summarize requirements and verify compliance.

5.3 Other Directions

Based on our results, we observe several other avenues of future work. First, we plan to investigate other data sources to further explore GDPR compliance in open-source projects. For example, we plan to mine relevant queries from Stack Overflow to gain insight into challenges and information needs developers have for implementing GDPR policies. We will also examine answers to observe how developers respond. For instance, online discussions between developers regarding policies often use disclaimers, such as the acronyms "IANAL" or "NAL" to indicate "I am not a lawyer", before offering advice or answering questions related to legal frameworks. Without legal expertise, we anticipate it is difficult for OSS developers to offer guidance and seek help complying with data privacy regulations–motivating the need for novel approaches to support regulation adherence and compliance assessment.

Moreover, we aim to engage with policymakers to understand their perspectives on data privacy policies and the challenges developers face implementing them. We will collect qualitative insights from politicians and individuals with authority to develop policies to further explore methods to support the implementation of privacy laws. Finally, we aim to extend this work to investigate the impact of broader technology-related policies on open-source software development practices–for instance, investigating the impact of alternative data privacy regulations (*i.e.,* the CCPA or CDPA) as well as other legal frameworks that will impact software development and maintenance, such as current and imminent legislation regarding artificial intelligence governance.

6 Related Work

We note two lines of related work: characterizations of stakeholder perspectives on data privacy regulations, and technical and methodological approaches for regulatory compliance.

Stakeholder perspectives: Research has investigated perspectives on the GDPR for stakeholders in data privacy regulation compliance. Sirur and colleagues examined organizational perceptions on the feasibility of implementing GDPR concepts, finding that larger organizations were confident in their ability to comply while smaller companies struggled with the breadth and ambiguity in GDPR requirements [93]. Earp *et al.* surveyed software users to show the Internet privacy protection goals and policies for online websites do not meet users' expectations for privacy [41]. Similarly, Strycharz *et al.* surveyed consumers to uncover frustrations and negative attitudes related to the GDPR [99]. Our work focuses on the perceptions of developers, who are responsible for implementing code changes to comply with data privacy regulations.

On the perspective of software engineers as regulatory stakeholders, van Dijk and colleagues provide an overview of the transition of privacy policies from self-imposed guidelines from developers to legal frameworks and legislation [108]. Alhazmi interviewed software developers to uncover barriers for adopting GDPR principles–finding the lack of familiarity, precedented techniques, useful help resources, and prioritization from employers. The paper also found that developers generally do not prioritize privacy features in their projects, focusing instead on functional requirements prevent compliance [20]. Similarly, researchers interviewed senior engineers to understand the challenges implementing general privacy guidelines, indicating a frustration with legal interactions and the non-technical aspects of requirements [29]. Finally, Klymenko *et al.* interviewed technical and legal professionals to investigate measures for data privacy compliance in GDPR implementation—noting a lack of understanding and need for interdisciplinary solutions [65]. 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 OSS developers.

Implementing and verifying GDPR compliance: Prior work has explored approaches to implement and verify GDPR compliance. For instance, Martín *et al.* recommend Privacy by Design methods and tools for GDPR compliance [77]. Shastri and colleagues introduce GDPRBench, a tool to assess the GDPR compliance of databases [91]. Li *et al.* investigated automated GDPR compliance as part of continuous integration workflows [73]. Al-Slais conducted a literature review to develop a taxonomy privacy implementation approaches to guide GDPR compliance [19]. Finally, Mahindrakar *et al.* proposed the use of blockchain technologies to validate personal data compliance [76]. Rather than proposing new software engineering methods, measures, and tools related to GDPR, our work takes an empirical perspective to understand current practices.

7 Threats to Validity

We discuss three types of threats to validity.

Construct: In mining OSS repositories, we defined the construct of "GDPR-related pull requests" based on the presence of the string "GDPR". Some PRs may incorrectly refer to GDPR (false positives), while others may perform GDPR-relevant changes without using the acronym (false negatives). Further, this construct does not provide proof of developers' knowledge or experience with GDPR policies. This is also biased towards English-speakers, as the GDPR acronym differs in other languages. To mitigate non-English GDPR-related PRs polluting the non-GDPR-related dataset, we manually inspected PR titles for various iterations of the GDPR in other languages, including "RGPD" (French, Spanish, and Italian), "DSGVO" (German), and "AVG" (Dutch). Moreover, neglecting other international interpretations may have also affected our sentiment analysis results. We used off-the-shelf NLP techniques to assess sentiment, inheriting biases from these methods (*e.g.,* misinterpreted connotations of homonyms such as "mock"). In addition, parametric models for sentiment analysis are based on defined dictionary values and cannot detect certain aspects of human communication, such as sarcasm. Prior work also suggests sentiment analysis tools can be inaccurate in software engineering contexts [63]. However, we use this to gain preliminary insights into developers' perceptions of GDPR compliance in OSS.

Internal: We perceive no internal threats. This study provides characterizations rather than cause-effect measurements.

External: There are several threats to the generalizability of our findings. We inherit the standard perils of mining OSS [64]. We focus on OSS available on GitHub, which omits other code hosting platforms, such as GitLab, which may be used by different populations of developers. We doubt our results generalize to commercial software, since those development organizations directly face the consequences of GDPR non-compliance. We only consider the effect of GDPR because it is the most prominent privacy law, and hence has the most available data. Other regulations may have different effects. Specifically, we conjecture differences in the software engineering impact between general data privacy regulations, such as the GDPR and CCPA, and industry-specific data privacy regulations, such as HIPAA and FERPA: general regulations may necessarily be more ambiguous.

8 Conclusions

Data privacy regulations are being introduced to prevent data controllers from misusing users' information and to protect individuals. To adhere with these regulations, developers are charged with the complex task of understanding policies and making modifications to the source code of applications to implement privacy-related requirements. This work examines the impact of data privacy regulations on software development processes by investigating code contributions and developer perceptions of GDPR compliance in OSS. Our results show that complying with data privacy regulations significantly impacts development activities on GitHub, evoking negative perceptions and frustrations from developers. Our findings provide implications for developers and policymakers to support the implementation of data privacy regulations that protect the rights of human users in digital environments.

9 Data Availability

We have uploaded the survey, datasets, codebook, and data collection and analysis scripts as supplementary materials [2]. Our IRB protocol does not allow us to share individual survey responses.

Acknowledgments

Brown and Brantly acknowledge support from the Virginia Commonwealth Cyber Initiative (CCI).

References

[1] [n. d.]. https://edps.europa.eu/data-protection/data-protection/legislation/history-general-data-protection-regulation_en
[2] [n. d.]. https://github.com/code-world-no-blanket/ESEM24-GDPR-OSS-Impact
[3] [n. d.]. MIT License. https://opensource.org/licenses/MIT. Accessed: July 2023.
[4] [n. d.]. Right to erasure ('right to be forgotten'). https://gdpr-info.eu/art-17-gdpr/.
[5] 1974. Family Educational Rights and Privacy Act of 1974. 20 U.S.C. § 1232g; 34 CFR Part 99. https://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html
[6] 1991. GNU General Public License, version 2. Free Software Foundation. https://www.gnu.org/licenses/old-licenses/gpl-2.0.en.html
[7] 1996. Health Insurance Portability and Accountability Act of 1996. Pub. L. No. 104-191, 110 Stat. 1936. https://www.govinfo.gov/content/pkg/PLAW-104publ191/pdf/PLAW-104publ191.pdf
[8] 2004. Apache License, Version 2.0. Apache Software Foundation. https://www.apache.org/licenses/LICENSE-2.0
[9] 2010. IEC 61508-1:2010 - Functional safety of electrical/electronic/programmable electronic safety-related systems – Part 1: General requirements. International Electrotechnical Commission. https://webstore.iec.ch/publication/5512
[10] 2014. ISO 90003:2014 - Software engineering – Guidelines for the application of ISO 9001:2015 to computer software. International Organization for Standardization. https://www.iso.org/standard/59149.html
[11] 2015. ISO 9001:2015 - Quality management systems – Requirements. International Organization for Standardization. https://www.iso.org/standard/62085.html
[12] 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/EC (General Data Protection Regulation). Official Journal of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679
[13] 2018. ISO 26262-1:2018 - Road vehicles – Functional safety – Part 1: Vocabulary. International Organization for Standardization. https://www.iso.org/standard/68383.html
[14] 2023. 5th State of CCPA & GDPR Privacy Rights Compliance Research Report – Q4 2022. Cytrio. https://cytrio.com/wp-content/uploads/2023/02/5th-State-of-CCPA-GDPR-Compliance-Report_FNL2.pdf
[15] 2023. GDPR Enforcement Tracker — list of GDPR fines. Enforcement Tracker. https://www.enforcementtracker.com
[16] Ahmad Abdellatif, Mairieli Wessel, Igor Steinmacher, et al. 2022. BotHunter: an approach to detect software bots in GitHub. 6–17.
[17] Abdel-Jaouad Aberkane, Geert Poels, and Seppe Vanden Broucke. 2021. Exploring automated gdpr-compliance in requirements engineering: A systematic mapping study. *Ieee Access* 9 (2021), 66542–66559.
[18] Saeed Akhlaghpour, Farkhondeh Hassandoust, et al. 2021. Learning from enforcement cases to manage gdpr risks. *MIS Quarterly Executive* 20, 3 (2021).
[19] Yaqoob Al-Slais. 2020. Privacy Engineering Methodologies: A survey. In *2020 International Conference on Innovation and Intelligence for Informatics, Computing and Technologies (3ICT)*. 1–6.
[20] Abdulrahman Alhazmi and Nalin Asanka Arachchilage. 2021. I'm all ears! listening to software developers on putting GDPR principles into software development practice. *Personal and Ubiquitous Computing* 25, 5 (2021), 879–892.
[21] Keri Allan. 2007. Reskilling for compliance. *Info. Professional* 4, 1 (2007), 20–23.
[22] Fernando Almeida and José Augusto Monteiro. 2021. Exploring the effects of GDPR on the user experience. *Journal of information systems engineering and management* 6, 3 (2021).
[23] Murugan Anandarajan, Chelsey Hill, Thomas Nolan, Murugan Anandarajan, Chelsey Hill, and Thomas Nolan. 2019. Text preprocessing. *Practical text analytics: Maximizing the value of text data* (2019), 45–59.
[24] Maythee Anegboonlap. 2018. Will this conflict with GDPR? https://github.com/ReferralCandy/woocommerce-referralcandy/pull/24/#discussion_r238153546. GitHub repository: ReferralCandy/woocommerce-referralcandy.
[25] Pauline Anthonysamy, Awais Rashid, and Ruzanna Chitchyan. 2017. Privacy requirements: present & future. In *International Conference on Software Engineering: Software Engineering in Society*. IEEE.
[26] Johannes Bader, Jonathan Aldrich, and Éric Tanter. 2018. Gradual program verification. In *Verification, Model Checking, and Abstract Interpretation*. Springer.
[27] Ben Balter. 2015. Open source license usage on GitHub.com. GitHub Blog. https://github.blog/2015-03-09-open-source-license-usage-on-github-com/
[28] Muneera Bano. 2015. Addressing the challenges of requirements ambiguity: A review of empirical literature. In *International Workshop on Empirical Requirements Engineering*. IEEE.
[29] Kathrin Bednar, Sarah Spiekermann, and Marc Langheinrich. 2019. Engineering Privacy by Design: Are engineers ready to live up to the challenge? *The Information Society* 35, 3 (2019), 122–142.
[30] Yoav Benjamini and Yosef Hochberg. 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 (1995), 289–300.
[31] Ani Betts. 2021. Just enough EULA to not get banned. https://github.com/anaisbetts/sirene/pull/37. GitHub repository: anaisbetts/sirene.
[32] Alex Bowyer, Jack Holt, Josephine Go Jefferies, Rob Wilson, David Kirk, and Jan David Smeddinck. 2022. Human-GDPR interaction: Practical experiences of accessing personal data. 1–19.
[33] Randolph E Bucklin and Catarina Sismeiro. 2009. Click here for Internet insight: Advances in clickstream data analysis in marketing. *Journal of Interactive marketing* 23, 1 (2009), 35–48.
[34] Noel Carroll and Ita Richardson. 2016. Software-as-a-medical device: demystifying connected health regulations. *Journal of Systems and Information Technology* 18, 2 (2016), 186–215.
[35] Ann Cavoukian. 2009. Privacy by design. (2009).
[36] David Chisnall. 2012. *The Go programming language phrasebook*. Addison-Wesley.
[37] Bernard CK Choi, Tikki Pang, Vivian Lin, et al. 2005. Can scientists and policy makers work together? *Journal of Epidemiology & Community Health* 59, 8 (2005), 632–637.
[38] Tom Clancy. 1995. The chaos report. *The Standish Group* (1995).
[39] Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. Routledge.
[40] Jose Luis de La Vara, Markus Borg, Krzysztof Wnuk, and Leon Moonen. 2016. An industrial survey of safety evidence change impact analysis practice. *IEEE Transactions on Software Engineering* 42, 12 (2016), 1095–1117.
[41] J.B. Earp, A.I. Anton, L. Aiman-Smith, and W.H. Stufflebeam. 2005. Examining Internet privacy policies within the context of user privacy values. *IEEE Transactions on Engineering Management* 52, 2 (2005), 227–237.
[42] Pietro Ferrara, Nicola Fausto Spoto, et al. 2018. Static analysis for GDPR compliance. In *CEUR Workshop Proceedings*. CEUR Workshop Proceedings, 1–10.
[43] Aaron J Fischer, Brandon K Schultz, Melissa A Collier-Meek, et al. 2018. A critical review of videoconferencing software to support school consultation. *International Journal of School & Educational Psychology* 6, 1 (2018), 12–22.
[44] Lucas Franke, Huayu Liang, Aaron Brantly, James C. Davis, and Chris Brown. 2024. A First Look at the General Data Protection Regulation (GDPR) in Open-Source Software. In *International Conf on Software Engineering: Companion Proceedings* (Lisbon, Portugal) *(ICSE-Companion '24)*. Association for Computing Machinery, New York, NY, USA, 2 pages.
[45] GDPR. 2018. Art. 4 GDPR: Definitions. https://gdpr.eu/article-4-definitions/
[46] GDPR. 2018. Art. 83 GDPR: General conditions for imposing administrative fines. https://gdpr.eu/article-83-conditions-for-imposing-administrative-fines/
[47] GitHub. 2022. Octoverse 2022: The state of open source software. https://octoverse.github.com
[48] Georgios Gousios and Andy Zaidman. 2014. A dataset for pull-based development research. In *Conference on Mining Software Repositories*. 368–371.
[49] Emitza Guzman, David Azócar, and Yang Li. 2014. Sentiment analysis of commit comments in GitHub: an empirical study. In *Mining Software Repositories (MSR)*.
[50] Rajaa El Hamdani et al. 2021. A combined rule-based and machine learning approach for automated GDPR compliance checking. In *International Conference on Artificial Intelligence and Law*.
[51] Nikolay Harutyunyan. 2020. Managing your open source supply chain-why and how? *Computer* 53, 6 (2020), 77–81.
[52] Paul Hitlin, Rainie Lee, and Kenneth Olmstead. 2019. Facebook Algorithms and Personal Data. Pew Research Center. https://www.pewresearch.org/internet/2019/01/16/facebook-algorithms-and-personal-data/
[53] Chris Hobbs. 2019. *Embedded software development for safety-critical systems*. CRC Press.
[54] Sebastian Holst. 2017. GDPR liability: software development and the new law. *LinkedIn* (2017). https://www.linkedin.com/pulse/gdpr-liability-software-development-new-law-sebastian-holst/
[55] Minqing Hu and Bing Liu. 2004. Mining opinion features in customer reviews. In *AAAI*, Vol. 4. 755–760.
[56] Syed Fatiul Huq, Ali Zafar Sadiq, and Kazi Sakib. 2019. Understanding the effect of developer sentiment on fix-inducing changes: An exploratory study on github pull requests. In *Asia-Pacific Software Engineering Conference*. IEEE.
[57] Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *AAAI conf on web and social media*.
[58] International Association of Privacy Professionals. Accessed 2023. *Global Comprehensive Privacy Law Mapping Chart*. https://iapp.org/resources/article/global-comprehensive-privacy-law-mapping-chart/
[59] International Electrotechnical Commission. 2010. Functional safety of electrical/electronic/programmable electronic safety-related systems - Part 3: Software requirements. https://webstore.iec.ch/publication/9277
[60] Georgeta Ion, Mihaela Stîngu, and Elena Marin. 2019. How can researchers facilitate the utilisation of research by policy-makers and practitioners in education? *Research Papers in Education* 34, 4 (2019), 483–498.
[61] Jim Isaak and Mina J. Hanna. 2018. User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection. *Computer* 51, 8 (2018), 56–59.

[62] Arthur M Jacobs. 2019. Sentiment analysis for words and fiction characters from the perspective of computational (neuro-) poetics. *Frontiers in Robotics and AI* 6 (2019), 53.

[63] Robbert Jongeling, Proshanta Sarkar, Subhajit Datta, and Alexander Serebrenik. 2017. On negative results when using sentiment analysis tools for software engineering research. *Empirical Software Engineering* 22 (2017), 2543–2584.

[64] Eirini Kalliamvakou, Georgios Gousios, Kelly Blincoe, Leif Singer, Daniel M German, and Daniela Damian. 2014. The promises and perils of mining github. In *11th Working Conference on Mining Software Repositories (MSR)*. 92–101.

[65] Oleksandra Klymenko, Oleksandr Kosenkov, Stephen Meisenbacher, et al. 2022. Understanding the implementation of technical measures in the process of data privacy compliance: A qualitative study. In *International Symposium on Empirical Software Engineering and Measurement*.

[66] Michael Kretschmer, Jan Pennekamp, and Klaus Wehrle. 2021. Cookie banners and privacy policies: measuring the impact of the gdpr on the web. *ACM Transactions on the Web)* 15, 4 (2021), 1–42.

[67] Oksana Kulyk, Nina Gerber, Annika Hilt, et al. 2020. has the gdpr hype affected users' reaction to cookie disclaimers? *Journal of Cybersecurity* 6, 1 (2020).

[68] Aman Kumar, Manish Khare, and Saurabh Tiwari. 2022. Sentiment Analysis of Developers' Comments on GitHub Repository: A Study. In *International Conference on Advanced Computational Intelligence (ICACI)*. IEEE, 91–98.

[69] Christian Kurtz, Martin Semmann, and Tilo Böhmann. 2018. Privacy by design to comply with GDPR: a review on third-party data processors. (2018).

[70] Daniël Lakens. 2013. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in psychology* 4 (2013), 62627.

[71] Roslyn Layton and Silvia Elaluf-Calderwood. 2019. A social economic analysis of the impact of GDPR on security and privacy practices. In *Conference on Cybersecurity and Privacy*. IEEE, 1–6.

[72] He Li, Lu Yu, and Wu He. 2019. The impact of GDPR on global technology development. *Journal of Global Information Technology Management* 22, 1 (2019).

[73] Ze Shi Li, Colin Werner, and Neil Ernst. 2019. Continuous Requirements: An Example Using GDPR. In *International Requirements Engineering Conference Workshops (REW)*. 144–149.

[74] MH Lloyd and PJ Reeve. 2009. IEC 61508 and IEC 61511 assessments-some lessons learned. (2009).

[75] Thomas W MacFarland, Jan M Yates, Thomas W MacFarland, and Jan M Yates. 2016. Mann–whitney u test. *Introduction to nonparametric statistics for the biological sciences using R* (2016), 103–132.

[76] Abhishek Mahindrakar and Karuna Pande Joshi. 2020. Automating GDPR Compliance using Policy Integrated Blockchain. 86–93.

[77] Yod-Samuel Martin and Antonio Kung. 2018. Methods and tools for GDPR compliance through privacy and data protection engineering. In *IEEE European Symposium on Security and Privacy–Workshops*. IEEE, 108–111.

[78] J. M. Valdez Mendia and J. J. A. Flores-Cuautle. 2022. Toward customer hyper-personalization experience — A data-driven approach. *Cogent Business & Management* 9, 1 (2022), 2041384.

[79] Dan Milmo and Lisa O'Carroll. 2023. Facebook owner Meta fined €1.2bn for mishandling user information. The Guardian. https://www.theguardian.com/technology/2023/may/22/facebook-fined-mishandling-user-information-ireland-eu-meta

[80] Rene Moquin and Robin L Wakefield. 2016. The roles of awareness, sanctions, and ethics in software compliance. *Journal of Computer Info. Sys.* 56, 3 (2016).

[81] Frank Nagle, James Dana, Jennifer Hoffman, Steven Randazzo, and Yanuo Zhou. 2022. Census II of Free and Open Source Software—Application Libraries. *Linux Foundation, Harvard Laboratory for Innovation Science (LISH) and Open Source Security Foundation (OpenSSF)* 80 (2022).

[82] Chinenye Okafor et al. 2022. Sok: Analysis of software supply chain security by establishing secure design properties. In *ACM SCORED Workshop*. 15–24.

[83] Kang-il Park and Bonita Sharif. 2021. Assessing perceived sentiment in pull requests with emoji: evidence from tools and developer eye movements. In *International Workshop on Emotion Awareness in Software Engineering*. IEEE, 1–6.

[84] Luca Pascarella, Davide Spadini, et al. 2018. Information needs in contemporary code review. *Proc. of the ACM on Human-Computer Interaction: CSCW* (2018).

[85] Cole S Peterson, Jonathan A Saddler, Natalie M Halavick, and Bonita Sharif. 2019. A gaze-based exploratory study on the information seeking behavior of developers on stack overflow. In *CHI*. 1–6.

[86] Daniel Pletea, Bogdan Vasilescu, and Alexander Serebrenik. 2014. Security and emotion: sentiment analysis of security discussions on github. In *Conf on mining software repositories*.

[87] Pricewaterhouse Coopers. 2017. Pulse survey: US companies ramping up general data protection regulation (GDPR) budgets. https://www.pwc.com/us/en/services/consulting/library/gdpr-readiness.html

[88] Martin Rinard. 2007. Automated techniques for surviving (otherwise) fatal software errors. *Electronic Notes in Theoretical Computer Science* 174, 4 (2007).

[89] Jane Ritchie and Liz Spencer. 2002. Qualitative data analysis for applied policy research. In *Analyzing qualitative data*. Routledge, 173–194.

[90] Adithya Sethi. 2021. Avoid lawsuits by mentioning cookies thing. https://github.com/ShizukuIchi/winXP/pull/100. GitHub repository: ShizukuIchi/winXP.

[91] Supreeth Shastri et al. 2020. Understanding and benchmarking the impact of GDPR on database systems. *VLDB* 13, 7 (2020), 1064–1077.

[92] Jeremy Siek and Walid Taha. 2007. Gradual typing for objects. In *European Conference on Object-Oriented Programming*. Springer, 2–27.

[93] Sean Sirur, Jason R.C. Nurse, and Helena Webb. 2018. Are We There Yet? Understanding the Challenges Faced in Complying with the General Data Protection Regulation (GDPR). In *Workshop on Multimedia Privacy and Security*.

[94] Ian Sommerville. 2011. *Software Engineering, 9/E*. Pearson Education India.

[95] Jeff South. 2018. More than 1,000 U.S. news sites are still unavailable in Europe, two months after GDPR took effect. Nieman Lab. https://www.niemanlab.org/2018/08/more-than-1000-u-s-news-sites-are-still-unavailable-in-europe-two-months-after-gdpr-took-effect/

[96] Richard Sproat, Alan W Black, Stanley Chen, et al. 2001. Normalization of non-standard words. *Computer speech & language* 15, 3 (2001), 287–333.

[97] David Stokes. 2012. 21 - Validation and regulatory compliance of free/open source software. In *Open Source Software in Life Science Research*, Lee Harland and Mark Forster (Eds.). Woodhead Publishing, 481–504.

[98] Margaret-Anne Storey, Neil A Ernst, Courtney Williams, and Eirini Kalliamvakou. 2020. The who, what, how of software engineering research: a socio-technical framework. *Empirical Software Engineering* 25 (2020).

[99] Joanna Strycharz, Jef Ausloos, and Natali Helberger. 2020. Data protection or data frustration? Individual perceptions and attitudes towards the GDPR. *Eur. Data Prot. L. Rev.* 6 (2020), 407.

[100] Synopsys. 2023. Open Source Security and Risk Analysis Report. https://www.pwc.com/us/en/services/consulting/library/gdpr-readiness.html

[101] Mohammad Tahaei, Tianshi Li, and Kami Vaniea. 2022. Understanding privacy-related advice on stack overflow. *Proceedings on Privacy Enhancing Technologies* (2022).

[102] Mohammad Tahaei, Kami Vaniea, and Naomi Saphra. 2020. Understanding privacy-related questions on stack overflow. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–14.

[103] Aurelia Tamò-Larrieux and Aurelia Tamò-Larrieux. 2018. Privacy by Design for the Internet of Things: A Startup Scenario. *Designing for Privacy and its Legal Framework: Data Protection by Design and Default for the Internet of Things* (2018), 203–226.

[104] Neil Thurman. 2020. Many EU visitors shut out of US sites in response to GDPR never came back. Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/news/many-eu-visitors-shut-out-us-sites-response-gdpr-never-came-back

[105] Serj Tubin. 2023. GDPR stuff. https://github.com/2beens/serj-tubin-vue/pull/71. GitHub repository: 2beens/serj-tubin-vue.

[106] UNCTAD. 2021. Data Protection and Privacy Legislation Worldwide. *United Nations Conference on Trade and Development* (2021). https://unctad.org/page/data-protection-and-privacy-legislation-worldwide

[107] Christine Utz, Martin Degeling, Sascha Fahl, et al. 2019. (Un) informed consent: Studying GDPR consent notices in the field. In *Conference on Computer and Communications Security*.

[108] N. van Dijk, A. Tanas, K. Rommetveit, and C. Raab. 2018. Right engineering? the redesign of privacy and Personal Data Protection. *International Review of Law, Computers & Technology* 32, 2–3 (Apr 2018), 230–256.

[109] Ana Vazão, Leonel Santos, Maria Beatriz Piedade, and Carlos Rabadão. 2019. SIEM open source solutions: a comparative study. In *2019 14th Iberian Conference on Information Systems and Technologies (CISTI)*. IEEE, 1–5.

[110] Denis Verdon. 2006. Security policies and the software developer. *IEEE Security & Privacy* 4, 4 (2006), 42–49.

[111] Branka Vuleta. 2023. 10 unbelievable GDPR statistics in 2023. https://legaljobs.io/blog/gdpr-statistics/

[112] Yue Wang, Irene L Manotas Gutièrrez, Kristina Winbladh, and Hui Fang. 2013. Automatic detection of ambiguous terminology for software requirements. In *18th International Conference on Applications of Natural Language to Information Systems (NLDB)*. Springer, 25–37.

[113] R Kent Weaver. 2015. Getting people to behave: Research lessons for policy makers. *Public Administration Review* 75, 6 (2015), 806–816.

[114] Krzysztof Wnuk, Tony Gorschek, and Showayb Zahda. 2013. Obsolete software requirements. *Information and Software Technology* 55, 6 (2013), 921–940.

[115] Christopher Wylie. 2019. How I Helped Hack Democracy. New York Magazine. https://nymag.com/intelligencer/2019/10/book-excerpt-mindf-ck-by-christopher-wylie.html

[116] Christopher Wylie. 2019. I Made Steve Bannon's Psychological Warfare Tool: Meet the Cambridge Analytica Whistle-blower. New York Magazine. https://nymag.com/intelligencer/2019/10/book-excerpt-mindf-ck-by-christopher-wylie.html

[117] Haoxiang Zhang, Shaowei Wang, Tse-Hsun Chen, Ying Zou, and Ahmed E Hassan. 2019. An empirical study of obsolete answers on stack overflow. *IEEE Transactions on Software Engineering* 47, 4 (2019), 850–862.