AN EMPIRICAL STUDY OF TRUST & SAFETY ENGINEERING IN OPEN-SOURCE SOCIAL MEDIA PLATFORMS

by

Geoffrey Cramer

A Thesis

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Master of Science in Electrical and Computer Engineering

School of Electrical and Computer Engineering

West Lafayette, Indiana

May 2023

THE PURDUE UNIVERSITY GRADUATE SCHOOL STATEMENT OF COMMITTEE APPROVAL

Dr. James C. Davis, Chair

Elmore Family School of Electrical and Computer Engineering

Dr. Alex Quinn

Elmore Family School of Electrical and Computer Engineering

Dr. Alice Marwick

Department of Communication

University of North Carolina at Chapel Hill

Approved by:

Dr. Dimitrios Peroulis

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Davis for his continued guidance and motivation throughout my work on this thesis. Without his support, this work would not have been possible. I would like to thank Dr. Marwick and Dr. Quinn for their consistent guidance with my work. I am indebted to my research colleagues, peers, and friends for their help and suggestions. I am especially thankful to my research partner, William Trey Maxam, for his significant contributions to this study. Finally, I do not take for granted the unwavering support that my fiancée, family, and parents have provided me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

T&S	Trust & Safety
SMP	Social Media Platform
OSS	Open Source Software
TSPA	Trust & Safety Professional Association
SDLC	Software Development Life Cycle

GLOSSARY

- **Diaspora**

  An open-source, decentralized SMP that is focused on social networking.

- **GitHub**

  A web-based platform that provides hosting for software development and version control.

- **GitHub Issues**

  A feature within the GitHub platform that allows users to track, manage, and discuss bugs, feature requests, and other issues related to the project.

- **Mastodon**

  An open-source, decentralized SMP that mimicks Twitter.

- **Open Source Software (OSS)**

  Software where the source code is freely available to the public, allowing anyone to view, modify, and distribute the software, often collaboratively and under an open-source license.

- **Risk management**

  Coordinated activities to direct and control an organization with regard to risk.

- **Social Media Platform (SMP)**

  Internet-based and persistent channels of mass personal communication facilitating perceptions of interactions among users, deriving value primarily from user-generated content.

- **Software Development Life Cycle (SDLC)**

  A structured process that outlines the steps involved in designing, creating, testing, deploying, and maintaining software.

- **Trust & Safety (T&S)**

  The study of how people abuse the internet to cause real human harm.

- **T&S Engineering**

  An area of software engineering focused on designing systems with T&S in mind.

- **T&S Professional Association (TSPA)**

  An association that supports the global community of professionals who develop and enforce principles, policies, and practices that define acceptable behavior and content online and/or facilitated by digital technologies.

- **T&S risk**

  The potential loss that users face when harmed by other users.

- **User T&S in SMPs**

  The study of how users harm other users on SMPs.

- **User T&S Engineering in SMPs**

  Software engineering methods that use knowledge of T&S to reduce harmful user-to-user interactions on SMPs.

ABSTRACT

Social Media Platforms (SMPs) are used by almost 60% of the global population. Along with the ubiquity of SMPs, there are increasing Trust & Safety (T&S) risks that expose users to spam, harassment, abuse, and other harmful content online. *T&S Engineering* is an emerging area of software engineering striving to mitigate these risks. This study provides the first step in understanding this form of software engineering.

This study examines how T&S Engineering is practiced by SMP engineers. I studied two open-source (OSS) SMPs, Mastodon and Diaspora, which comprise 89% of the 9.6 million OSS SMP accounts. I focused on the T&S design process by analyzing T&S discussions within 60 GitHub issues. I applied a T&S discussion model to taxonomize the T&S risks, T&S engineering patterns, and resolution rationales. I found that T&S issues persist throughout a platform's lifetime, they are difficult to resolve, and engineers favor reactive treatments. To integrate findings, I mapped T&S engineering patterns onto a general model of SMPs. My findings give T&S engineers a systematic understanding of their T&S risk treatment options. I conclude with future directions to study and improve T&S Engineering, spanning software design, decision-making, and validation.

1. INTRODUCTION

Social Media Platforms (SMPs) are used by almost 60% of the global population [1]. Benefits of SMPs can be general (*e.g.,* sharing information, expression, and entertainment [2]) as well as context-specific (*e.g.,* improved marketing for small businesses [3] and alleviation of students' test anxiety [4]). However, there are also many documented harms of SMPs, including cyberbullying [5], sexual harassment [6], and online radicalization [7]. Many SMPs rely on manual and automated moderation [8] to address these concerns, balancing competing requirements including discourse, preserving the platform's trustworthy reputation, and keeping users safe.

SMPs are thus at the epicenter of an emerging engineering discipline called *Trust & Safety (T&S) Engineering*. The Trust & Safety Journal defines T&S as "the study of how people abuse the Internet to cause real human harm" [9].[1] A software engineer on the Trust & Safety team at GitHub defines T&S Engineering as "software designed with user safety in mind" [10]. If we better understand how SMPs can be designed to promote trust and safety, we will help software engineers improve human interactions worldwide. Researchers have previously investigated SMP problems [11]–[13] and potential solutions [14]–[17] but no prior work describes how T&S engineering is actually practiced. Addressing this gap will allow researchers and practitioners to understand, improve, and standardize this discipline.

In my thesis, I describe the first empirical study of T&S Engineering in SMPs. My goal was to characterize the T&S engineering design process. In particular, I wanted to learn what T&S risks are identified in which SMP features, what solutions are explored, and what properties are prioritized in solutions. I analyzed 60 T&S-related issues from two open-source SMPs, Mastodon and Diaspora. To do this, I sampled T&S issues using keywords, mapped the T&S engineering design process onto a discussion model, and analyzed elements of this discussion model. I used a mix of open- and closed- coding to develop taxonomies for T&S risks, engineering patterns, and pattern selection rationals in SMPs. I used inter-rater agreement to validate results.

[1] ↑The Trust & Safety Journal was established in 2021 by academics at Standford to " to bring together rigorous trust and safety research, which is currently spread across many disciplines, and to spur new research in this field" [9].

This study has four primary findings: T&S issues remain persistent throughout the SMP lifetime, most T&S issues highlight design shortcomings instead of implementation errors, T&S issues are difficult to resolve or remain open (the average T&S issue resolution time is 147 days longer than the total average), SMP engineers primarily select remedial treatments instead of preventative ones, SMP engineers typically transfer T&S risk to moderators and users.

To summarize my thesis:

- I conduct the first study of T&S Engineering, providing novel insight into the field.
- I contribute a novel method to extract and analyze OSS discussions for a specific subject of interest.
- I taxonomize T&S risks and threat actors (Table 5.2), providing researchers and practitioners a useful starting point for future work on T&S risk mitigation.
- I taxonomize T&S Engineering patterns (Table 5.3) and the contexts under which they operate (Figure 5.3), giving prior grey literature [18] an empirical basis.
- I taxonomize T&S decision rationales (Table 5.4), adapting prior work to the T&S engineering context.
- I contribute a coded dataset of 60 T&S discussions in real-world software projects, providing a starting point for future work in the T&S Engineering space.

Significance: Trust & Safety Engineering is an emerging focus for software engineers whose systems facilitate human interaction. Social Media Platforms are the most prominent such systems. My work provides the first characterization of the T&S engineering design process for SMPs. My methodology demonstrates a novel analysis of risk-based decision-making in software engineering. I develop taxonomies for risks, treatment patterns, and decision rationales in T&S engineering discussions. Using these taxonomies, I make empirically-based recommendations for how T&S engineers can implement more trust and safety into SMPs.

Thesis Statement: A taxonomy of T&S risks, solutions, and decision rationales can mature T&S Engineering as a discipline and further its research and development.

2. BACKGROUND

To contextualize this study, I review SMP definitions and types, provide a history of T&S, and discuss a risk management view of T&S in SMPs.

2.1 Social Media Platforms: Definition & Types

I begin with discussing what SMPs are, what their makeup is, how they vary, and how pervasive they have become. Carr & Hayes define social media comprehensively: "Internet-based... and persistent channel[s] of mass personal communication facilitating perceptions of interactions among users, deriving value primarily from user-generated content" [19]. SMPs have also been broken down into core components by Smith: identity, presence, relationships, reputation, groups, conversations, and sharing [20]. SMPs can take different forms by varying how strongly each of these components are emphasized on the platform [21]. Beyond definitions, SMPs are among the most popular services on the Internet: over half of the top 20 most visited websites being an SMP as of May 2022 [22].

Many SMPs are operated by for-profit businesses, introducing potential conflicts between profit and safety. The Twitter's recent decision to disband its Trust & Safety group exemplifies these tensions [23]. OSS SMPs try to address this concern. OSS SMPs emerged in 2010 in projects such as Diaspora, pump.io, and GNU Social [24]. Additional platforms like Mastodon and Pleroma came out around 2016 [24]. Most OSS SMPs are *decentralized*. In a decentralized SMP, an administrator can deploy an *SMP instance* on a server for public or private use. Content can be shared across SMP instances through activity stream protocols [25], creating a "Fediverse" (federated universe) [26]. This approach to social networking has created a dichotomy of SMP architectures: *commercial & centralized* or *open-source & decentralized*.[1]

[1] ↑ It would be incomplete to not mention philosophical differences between commercial and OSS SMPs. Commercial SMPs are for-profit businesses while OSS SMPs are freely available software applications, often run by non-profits. Due to their orientation, commercial SMPs often abate user safety issues, especially when it has a perceived negative impact on the business [27].

Figure 2.1 provides a simplified model of SMPs, focused on how platform actors (software engineers, moderators) influence user interactions.[2] *User-generated interactions* are any action that other users can see. This term encompasses *user-generated content* and includes posted content, replies, likes, reactions, reshares, quotes, user mentions, private messages, and anything else that one user does to influence what another user is exposed to.

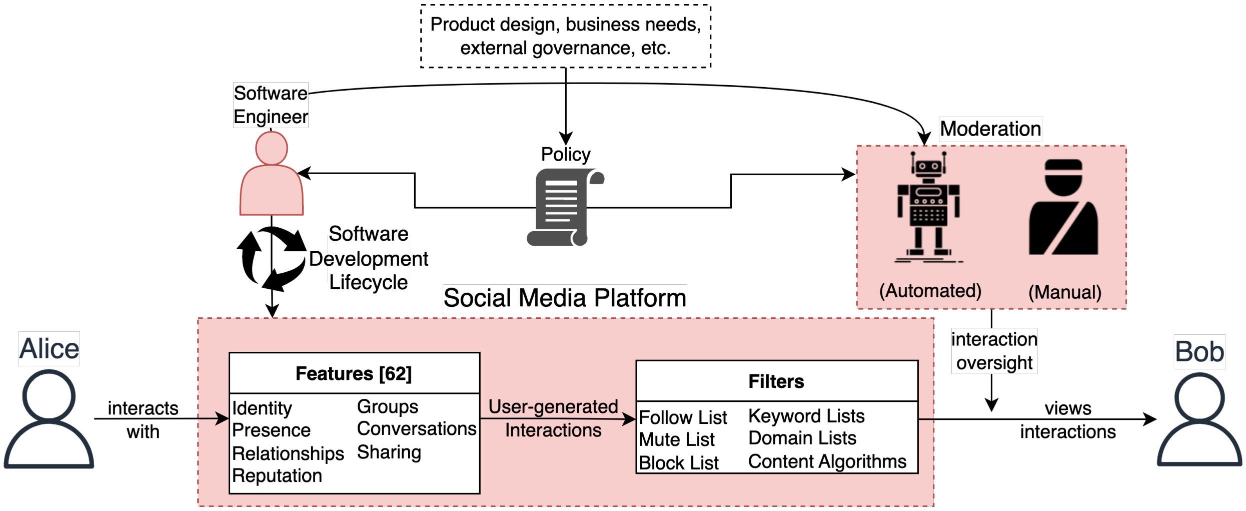

Figure 2.1. SMP context diagram showing a one-way interaction between Alice and Bob. Alice interacts with features and those interactions pass through filters and moderation oversight before reaching Bob. This study's focus is highlighted in pink. This study focuses on how software engineers design SMPs and moderation tools to influence how Alice and Bob can safely interact.

2.2 Trust & Safety and its Engineering

According to Cryst *et al.*, discussions of Trust & Safety originated in the financial sector in the 1990s to address issues such as fraudulent activity [9], [28]. Platform operators want users to *trust* the platform and feel *safe* on it, both in terms of their interactions with the platform provider (*e.g.*, not having their data exploited [29]) and in terms of their interactions with other users (*e.g.*, not being spammed or exposed to harmful content) [30]. Over time, it became clear that any digital platform where users interact will experience T&S issues. Efforts to promote T&S were initially distributed across teams, making it difficult

[2] ↑ The model was developed based on knowledge gained from background research and completing the study itself.

to consolidate best practices and apply research findings [31], [32]. These shortcomings prompted centralization: dedicated "Trust & Safety" teams charged with internal platform governance. Professionalization followed: the Trust and Safety Professional Association (TSPA) launched in 2020, with founding organizations including many SMPs (*e.g.,* Facebook, Twitter, Instagram, YouTube, and OKCupid) [32]. Concurrently, academics at Stanford founded the *Trust & Safety Journal* in 2021 [9].

Trust & Safety Engineering emerged as a discipline of software engineering in recent years. The goal of T&S Engineering is to consider T&S throughout the software development lifecycle, spanning requirements, design, implementation, validation, and operation (*e.g.,* moderation). I am not aware of prior academic literature that describes T&S Engineering. However, many companies employ T&S Engineers. GitHub says their T&S Engineers "design [software] with user safety in mind" [10] and Leong discusses community safety checks in GitHub release pipelines [33]. GitLab, Cloudflare, and Pinterest advertise T&S Engineering teams as well [34]–[37]. The TSPA job board lists many T&S opportunities calling for software engineering experience [38].

This study applies the concepts of T&S and T&S Engineering and reports the first examination of T&S Engineering in practice.

2.3 Trust & Safety in Social Media Platforms: A Risk Management View through ISO 31000:2018

T&S issues on SMPs are a global challenge. For example, a 2021 Pew Research Center survey of Americans found that ∼40% of respondents had experienced online harassment [6]. In 2022, as part of a United Nations action, several nations launched an effort to address online abuse such as on SMPs [39].

To scope the broad definition of T&S to my study of SMPs, I define: *User T&S in SMPs* as the study of how users harm other users on SMPs, and *User T&S Engineering in SMPs* as software engineering methods that use knowledge of T&S to reduce harmful user-to-user interactions on SMPs. I use "T&S in SMPs" as shorthand for both concepts, and let context distinguish them. This definition excludes T&S issues in the user-platform relationship, *e.g.,* issues about GDPR. There were relatively few such issues in the studied OSS SMPs, perhaps because OSS SMPs lack the profit motivation that leads some commercial platforms to violate T&S in this way. I omitted them during our sampling process.

Due to the uncertain nature of T&S, this study takes a risk-oriented approach and defines *T&S risk* as the potential loss that users face when harmed by other users. To organize prior research on T&S in SMPs, I apply the risk management framework from ISO 31000:2018 [40].[3] I focus specifically on the *risk assessment* and *risk treatment* stages of the framework. These are the stages that most directly involve engineers, and which are necessary even if other stages are omitted.[4] Other researchers have also described T&S challenges in SMPs using risk frameworks [41]–[43].

2.3.1 Risk Assessment

The *risk assessment* step spans the identification, analysis, and evaluation of risks, threats, and vulnerabilities.

Many sources have taxonomized T&S risks and threats on social media [11], [41], [44]–[46]. Hasib provided a foundational treatment of SMP risks, considering categories such as traditional information security (*e.g.,* spam, XSS), identity (*e.g.,* phishing, fake profiles), privacy (*e.g.,* digital dossiers, facial recognition), and social threats (*e.g.,* stalking) [44]. Laorden *et al.* used a threat modeling approach to SMPs to identify additional threats such as private information disclosure and corporate secrets theft [45]. Other researchers expanded these taxonomies, adding categories such as child-specific threats [11], privacy threats such as deanonymization and location leakage [41], [46], and political threats such as disinformation [47]. Thomas *et al.* provided the most recent and exhaustive taxonomy, enumerating myriad forms of online hate and harassment [13]: toxic content, content leakage, overloading, false reporting, impersonation, surveillance, and lockout & control.

Beyond taxonomies, researchers have investigated individual threats. For example: Trabelsi & Bouafif described abuses of content reporting systems [48]; Ashktorab and Vitak investigate cyberbullying mitigation and prevention techniques [5]; Usmani *et al.* analyze

[3]↑ ISO 31000:2018 is now behind a paywall. I summarize relevant content here.

[4]↑ The other stages are stakeholder communication, scoping, monitoring, and reporting. These stages are oriented toward engineering leadership and management, and could be omitted by some organizations.

social insider attacks [49]; Such *et al.* investigated privacy conflicts in co-owned photos [50]; and Cheng *et al.* studied the efforts of "trolls" to disrupt constructive discussion [51].

Due to its recency and sound methods, I view Thomas *et al.* [13] as the state-of-the-art taxonomy of T&S risks. I build on it, identifying two additional categories and extending a third.

2.3.2 Risk Treatment

In the *risk treatment* step, T&S engineers identify candidate treatments to mitigate risks. Two kinds of approaches are used to mitigate T&S risks on SMPs: design and moderation. Figure 2.1 illustrates these protection mechanisms.

Design Treatments are *proactive*, preventing T&S issues before they manifest by promoting *T&S By Design*. Unfortunately, there is a gap in literature that operationalizes how engineers can do this. Past work has investigated specific mechanisms to protect users such as improving authentication and user settings [11], using socially-aware content access control [52], [53], and experimenting with safe designs for specific interfaces [16], [54]. Related work from the more matured *privacy & security by design* [55] literature can advance safe design research, as this approach have been shown to prevent privacy failures in SMPs [56]. However, the closest work that attempts to generalize safe design are two pieces of grey literature:

- The *Safety By Design* framework [57] provides a process for the entirety of platform governance, with brief mentions of how to pursue safe design within software itself. Design strategies include: providing content reporting, communicating social contracts, implementing harmful content detection, practicing *privacy & security by design* [55], providing safety tools, leveraging technical features to mitigate risk, evaluating all features to mitigate risk factors, and publishing annual safety assessments.

- The taxonomy provided by Koscik [18] lists seven software design patterns to address online abuse vectors: remove feature, reduce interaction, reduce visibility, remove data, interaction intervention, require consent, and add moderation.

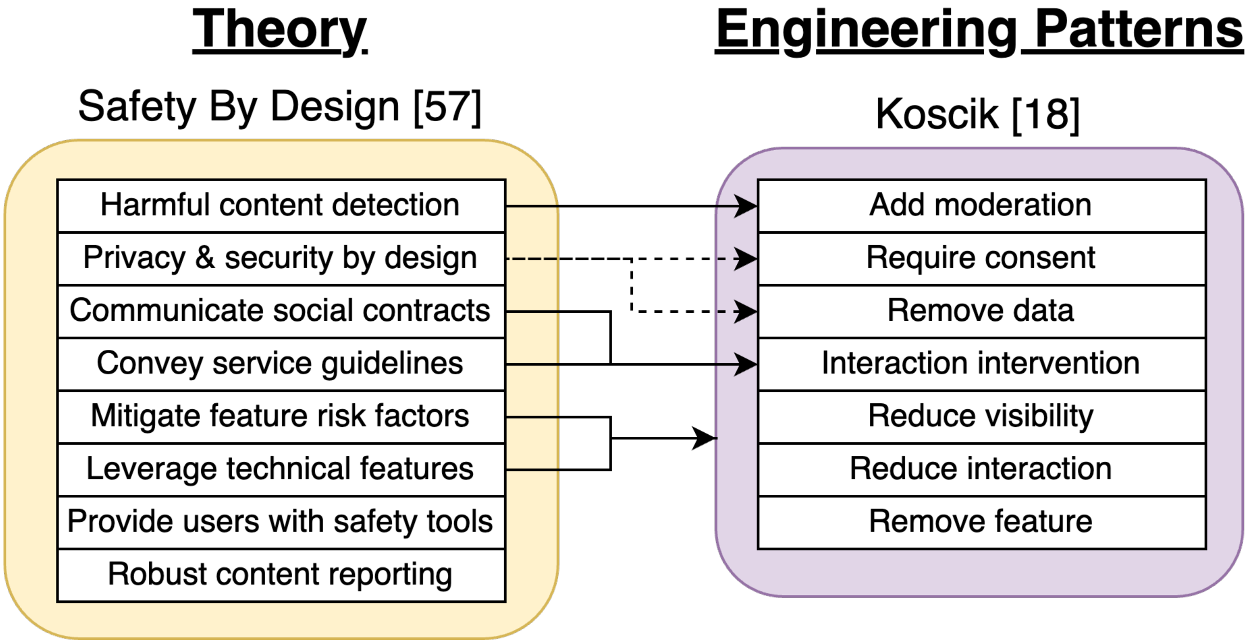

Figure 2.2. Existing literature that informs *T&S By Design*. The Safety By Design framework [57] on the left provides a theory-based set of guidelines to design safe platforms. Koscik [18] on the right lists seven patterns to address abuse vectors. Note that the Safety By Design list is not exhaustive and only lists items that involve technical work (for example, "develop community guidelines, terms of service and moderation procedures" was removed). Arrows indicate a relation and dashed arrows indicate a partial relation.

By studying design treatments more rigorously for T&S in SMPs, practitioners can develop measures that are effective [16], [54], scalable, and preventative. *T&S By Design* requires minimal engineering effort and operational cost but can prevent online harm from reaching targeted users and moderation staff.

Figure 2.2 shows both prior works and how they inform *T&S By Design*. I draw relations between elements if a suggestion from the *Safety By Design* framework can be implemented with a pattern from the abuse vector treatment taxonomy from Koscik. One of the full relations is "harmful content detection" and the *add moderation* pattern — per Koscik, the only way to detect harmful content with additional moderation. The only partial relation is "privacy & security by design" and the *require consent and remove data* patterns —

both patterns are realizations of the Privacy & Security By Design framework but do not encompass it.

Because both works are not rigorous nor based on empirical data, we do not know if they are exhaustive or if additional relations exist between them. This study aims to provide a solid foundation to these efforts, investigate new relationships that may exist, and gain novel insights into how T&S Engineering is practiced.

Moderation Treatments are *reactive*, limiting the impact of problematic user behaviors after they have occurred. For example, in Figure 2.1, moderation can only apply after Alice interacts with a feature, possibly before Bob sees her behavior. SMP moderation is carried out by platform administrators and automated systems. In many SMPs, moderation is manual, by volunteers or T&S teams [8]. Some platforms moderate automatically, including via per-user and community-based approaches [58].

Policies, both external and internal, may influence an SMP's approach to T&S. *External* policies such as the European Union's General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) promote T&S by regulating how organizations can access and process their users' personal information. Since SMPs derive value from user-generated content (§2.1), such policies affect SMP designs [59]. Additionally, many SMPs have *internal* platform policies developed by platform governance teams, including T&S teams [60]. These policies commonly describe acceptable user behavior (*e.g.,* codes of conduct) and may impact both system design and moderation.

Due to its novelty and comprehension, I view the grey literature abuse vector solution taxonomy of Koscik [18] as the most relevant work in the T&S Engineering field for addressing T&S risks. I build on it by identifying five additional categories to treat T&S risks.

2.3.3 Risk-based Decision Making

To select among candidate risk treatments, the ISO 31000:2018 standard outlines steps to perform risk-based decision-making. They are: (1) risk identification, (2) formulation of options, and (3) rationalization and selection of risk treatment plan. The standard indicates six general approaches for a risk treatment: eliminating the activity that gives rise to the risk; increasing the risk to pursue an opportunity; removing the source of the risk; changing the likelihood of the risk; changing its consequences; acknowledging but retaining the risk; and most notably for our study, *sharing the risk* among more parties so that each party faces less risk.

This risk-treatment-rationale model for decision-making is consistent with more general theories of argumentation used in the software engineering research literature [61], [62]. It permits us to build on the state of the art taxonomies for risks [13] and treatments [18] for T&S in SMPs. Since prior work has not considered T&S Engineering specifically, there is no specialized taxonomy for rationales. Among general software engineering rationale taxonomies, I found the rationale taxonomy of Al Safwan & Servant too fine-grained for this purpose [63], and instead contextualized the taxonomy developed by Ko *et al.* [64] (see Table 5.4 for results).

2.4 Summary and Unknowns

SMPs have a significant impact on society. Existing work takes a user-centric perspective in taxonomizing T&S threats in SMP threats, and an algorithmic view of treatments. We know little of the practice of T&S Engineering and risk-based T&S decision-making. By analyzing the T&S Engineering process in practice, the field can gain novel insight into how OSS T&S engineering decisions are made and how they might be improved.

3. RESEARCH QUESTIONS

Establishing effective T&S engineering practices for SMPs is critical to mitigating the widespread risks that have been discussed. This research provides initial steps toward achieving this goal by understanding the current practice. Specifically, the study analyzes: the context of T&S issues, the risk identification process, and the treatment selection process.

RQ1 *What are the characteristics of T&S issues?* To establish the contexts under which T&S issues arise, this study analyzes when they arise and in which features of the system.

RQ2 *What risks are identified in T&S issues?* This study determines which risks are identified and when, as well as the threat actors that manifest them.

RQ3 *What options are proposed in T&S issues? How are they selected?* Last, the study analyzes the proposed treatments, rationales for selecting them, and how effective the resolution process is.

4. METHODOLOGY

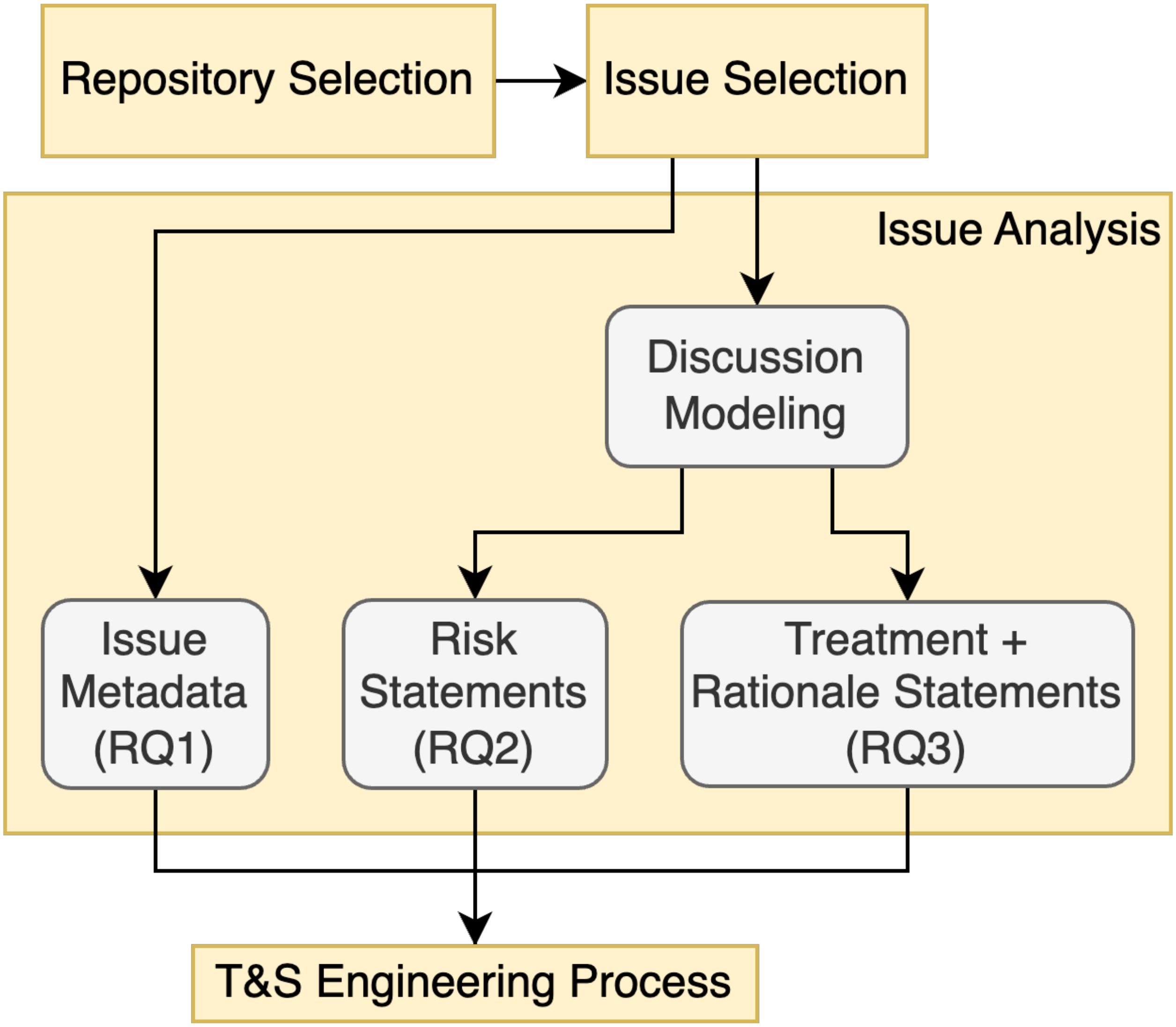

Figure 4.1. Relationship of research methods and data to RQs.

This study employs a *repository mining* method [65] to extract and analyze T&S discussions related to OSS SMPs. These repositories have thousands of issues (Table 4.1), many of which involve T&S topics such as privacy and harassment. Since this study is exploratory, mining repository data provides a cost-effective starting point to identify open challenges in T&S Engineering for future study. Figure 4.1 provides an overview of the methods for this study.

The mining approach has three steps. (1) OSS SMPs are selected. (2) T&S issues are identified via keywords. (3) T&S issues are analyzed. Specifically, I structured T&S issue dialogues following a discussion model (§4.3.2) and then coded the model elements for T&S themes and practices.

RQ1 is answered with T&S issue metadata. I determine: when they appear over time, which SMP features they occur in, which phase of the software development lifecycle (SDLC) they involve, and how long they take to resolve.

RQ2 is answered with T&S risk and threat actor taxonomies, based on the *risk* statements in the discussion model (§4.3.2).

RQ3 is answered with taxonomies of T&S engineering patterns, and T&S treatment rationales developed from *option* and *rationale* statements in the discussion model. Because *rationales* were only coded for closed T&S issues, I split the rationales based on the issue result of *merged* and *no action*.

4.1 Repository Selection

To select the specific OSS SMP projects for the study, I consulted an aggregated dataset of all such platforms [66] — see Table 4.1. Mastodon and Diaspora were chosen and were the top two by all counts.[1] The goal is to study T&S at scale and produce generalizable results. I chose to study the most popular repositories, a tested approach when mining data in software engineering research [68]–[71]. By selecting Mastodon and Diaspora, 89% of the OSS SMP user base can be studied and achieve this goal. Both projects use GitHub and track issues via "GitHub Issues" [72], [73].

4.2 Issue Selection

I used a keyword approach to find GitHub Issues containing T&S risk statements. Issue selection followed three phases: selecting baseline keywords, tailoring keywords to the studied projects, and sampling issues. An overview of the approach is provided in Figure 4.2.

[1]↑ Mastodon and Diaspora had 3.6M and 820K accounts respectively when data was pulled in April 2022. Shockingly, Mastodon's userbase has almost doubled since then, primarily due to Elon Musk's takeover of Twitter accompanied by significant platform governance changes [67].

Table 4.1. OSS SMP projects with over 100K accounts. The table shows accounts, GitHub issues, and GitHub stars as of January 26, 2023. Mastodon and Diaspora were chosen.

Project	Category	Active Accounts [66]	Issues	Stars
Mastodon	Microblogging	7,833,218	8,892	39.7K
Diaspora	Social networking	740,409	4,719	13.2K
PeerTube	Video sharing	288,964	4,386	11.4K
pixelfed	Photo sharing	150,326	1,702	4.5K
Pleroma	Microblogging	127,861	2,983	123
BirdsiteLive	Microblogging	101,188	91	398

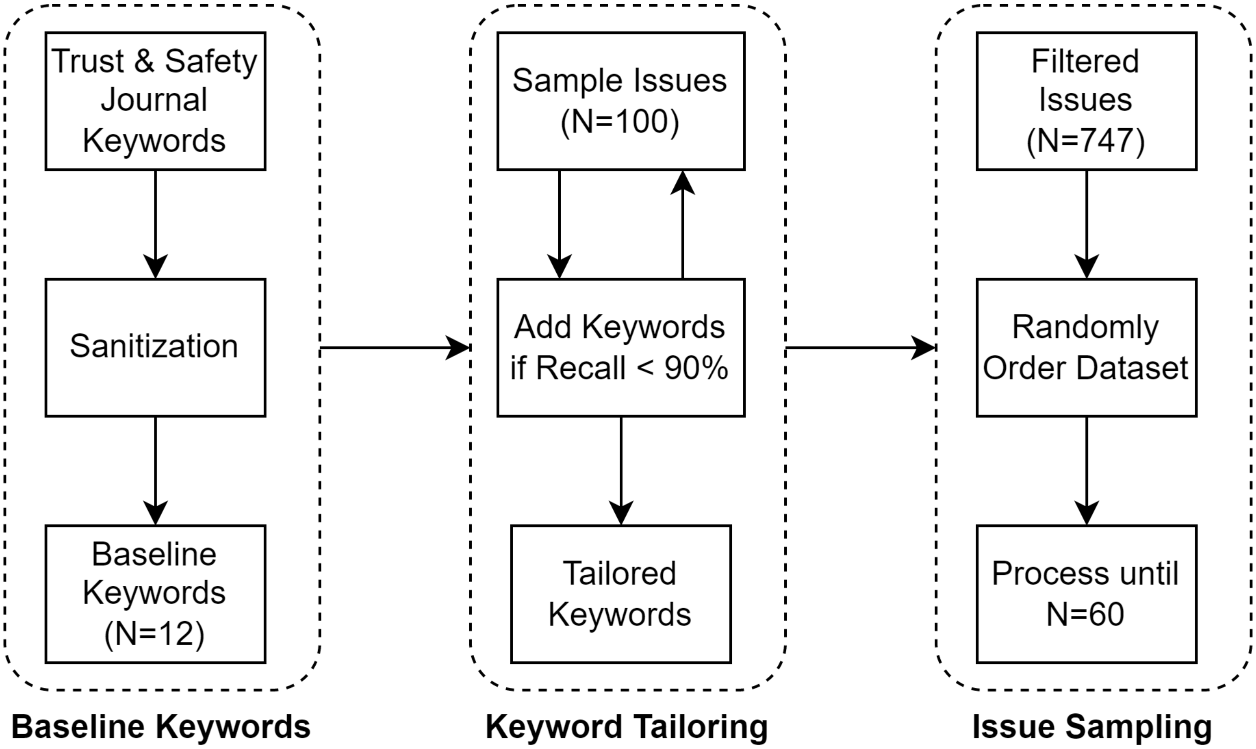

Figure 4.2. Overview of issue selection process.

4.2.1 Baseline keywords

I decided to filter issues based on keywords, a common approach when mining data in software engineering research [74], [75]. The baseline set of T&S keywords was formed by aggregating all keywords from the 15 articles in the first two issues of the *Trust & Safety Journal* [76]. 43 entries were removed if they were unrelated to the definition of *User T&S in SMPs* (*e.g.,* "robust hashing"), leaving 12 keywords.[2] Stemming and regular expressions were used to capture keyword variations. This step reduced Mastodon from 6,523 issues to 659 and Diaspora from 4,699 issues to 182. I applied an additional filter that issues should have at least 5 comments to ensure adequate discussion. This filter reduced Mastodon from 659 issues to 317 and Diaspora from 182 issues to 113.

4.2.2 Keyword Tailoring

Next, I tailored the keyword list to each selected repository. The goal was to find as many T&S discussions as possible. Rounds of 100 issues were randomly sampled on each of the two platforms. Additional keywords were added to each platform's keyword list in each round based on the *T&S in SMPs* definition (§2.3). I continued until the recall rate reached 90%. This cutoff choice was based primarily on intuition that the rate of false negatives should be as low as possible. This step expanded Mastodon from 317 issues to 431 and Diaspora from 113 issues to 316.

4.2.3 Issue Sampling

Finally, the issues that passed the filters and matched the keywords were randomly sorted for processing. I applied additional filters during this step: (1) Relevance based on the *T&S in SMPs* definition; (§2.3) and (2) discarding issues marked as duplicates. While processing issues, I found that issues with many comments were overwhelming to model (§4.3.2), so I filtered out issues with >20 comments (13 issues across both projects). I processed issues until a sample size of N=30 was reached in each repository (60 total). This stopping point

[2]↑ The baseline keywords are: moderation, suicide, self harm, fake news, misinformation, hate speech, harassment, governance, abuse, safety, cyberbullying, deepfakes.

was chosen due to resource constraints, but was sufficient to expand the state-of-the-art taxonomy in each dimension that was examined. A summary of the sampling process is given in Table 4.2.

Table 4.2. SMP filtering results, summarizing resulting keywords, precision and recall in the final batch of keyword expansion, number of T&S issues after the selection process, and proportion examined to reach 30 issues per project.

Project	Keywords	Prec., Rec.	Filtered Issues	Analysis %
Mastodon	17	50%, 100%	431	26%
Diaspora	15	27%, 100%	316	73%

4.3 Issue Analysis

After collecting issues, the unit of analysis was defined to be every sentence of every comment including the initial proposal. The resulting issues were analyzed as follows.

4.3.1 Issue Metadata

The following were labeled before issue discussions were analyzed:

- Issue Type: Issues were labeled as a *bug* the subject was about an error in implementation of the system and a *feature request* if the subject was about the design of the system.

- Issue Result: The result of the issue was labeled as *open* if it was still in the open state on GitHub. The issue could also be marked as *merged* if the engineers made some change to the system or *no action* if no change was made.

- Feature: The feature of the system was labeled based on the issue topic (see §4.3.3).

4.3.2 Discussion Modeling

Analyzing issue discussions is challenging due to their unstructured nature [61], [77], [78]. This study was focused on the T&S design process, so I modeled the discussions using the risk-based decision-making model described in §2.3.3. This model considers that an engineering decision requires treatment options, their associated risks, and rationales for choosing among them. I modeled each issue accordingly, discarding sentences that did not fall into any of these categories or were reiterations of previous points made in a discussion.

Risk I label *risk* identification statements if they contain a T&S risk claim (see §2.3). This category follows the Risk Identification step of the ISO 31000 standard [79].

Option I label *option* statements if they advance the issue towards closure (e.g. suggesting an implementation or proposing to take no action).

Treatment Selection Rationale I label *options* as *chosen* if they are accepted by developers and label the treatment selection *rationale* for accepting or rejecting them. I only coded *rationales* for *chosen options* because many unchosen ones did not have sufficient *rationale* claims, and because justifications for *chosen options* are most important.

Table 4.3 and Table 4.4 depict example issues.

4.3.3 Development of Taxonomies

A subsequent round of thematic coding was performed across issue and discussion model categories. Most taxonomies in this study leveraged existing literature to better contextualize the work. The base taxonomies are:

- The issue type was labeled as a *bug* if the subject was about an error in implementation of the system and a *feature request* if the subject was about the design of the system.

- The issue result was labeled as *open* if it was still in the open state on GitHub. The issue could also be marked as *merged* if the engineers made some change to the system or *no action* if no change was made.

- SMP feature list was developed from overarching issue topics and was openly coded (Table 5.1).

- T&S risk, threat actor taxonomies were developed from *risk* statements. The T&S risk taxonomy leveraged existing work from Thomas *et al.* [13] and was assigned based on the T&S risk definition (chapter 4). The threat actor taxonomy was developed from the

Table 4.3. Mastodon issue #9791 discusses a proposal to allow users to appeal moderator decisions (e.g. bans). ID numbers are sequential in time. Row 1 is the initial proposal. Row 2 claims row 1 would introduce a risk. Row 3 claims row 1 would treat a risk. Row 4 dismisses row 2 by saying it is inconsequential. Row 5 and row 6 add additional requirements. Row 7 claims row 1 would treat a risk. Row 1 and row 6 are chosen by engineers. In their solution, engineers added an appeal form and only allow it to be submitted once.

ID	User	Comment	Option	Risk	Rationale	Chosen
1	A	"form available to folks who are [banned] to be able to submit an appeal"	X		X	X
2	B	"will just be used as a method for bad actors to harass mods and admins"		X		
3	C	"[other sites have] trigger-happy mods [where] users have [been] abused"		X	X	
4	C	"Bad actors have enough means to get back at an admin if they want to"			X	
5	C	"make sure appeals go to other mods [or] it would encourage conflict"	X	X		
6	D	"the appeal can only happen once per a certain time limit"	X			X
7	E	"[current workaround] detaches the issue from the mod panel"		X	X	

Table 4.4. Diaspora issue #4664 asks for the ability to change the visibility of content after it has been posted. ID numbers are sequential in time. Row 1 is the initial proposal. Row 2 raises a risk of how the feature could be abused. Row 3 proposes an additional requirement to address the risk. Row 2 is chosen by engineers, indicating that the issue was closed with no action taken.

ID	User	Comment	Option	Risk	Rationale	Chosen
1	A	"Add ability to change a post scope after it's publication"	X			
2	B	"if someone comments your post thinking 'I can say what I want this is private' and then you change the visibility of the post, the comment becomes public too, so the whole internet has access to it."	X	X	X	X
3	B	"I was thinking of maybe allow to change visibility only if the post has no comment."	X			

basic user roles of OSS SMPs and extended when common actors were identified during coding. Assignment followed the threat actor definition (chapter 4).

- A T&S Engineering pattern taxonomy was developed from *option* statements that treat a T&S risk (Table 5.3). It extends work from Koscik *et al.* [18]. Annotators started with this taxonomy and iteratively developed new categories for statements that did not fit.

- A *rationale* taxonomy was developed from *rationale* statements (Table 5.4), leveraging existing work [64]. Specifically, I chose the *software quality* taxonomy from Ko *et al.* to capture the desired system properties that influenced decisions.

4.3.4 Inter-rater Agreement

Agreement was achieved between myself and another independent annotator. Agreement was measured using Cohen's Kappa coefficient [80]:

1. First, I (1) coded issue types and results (§4.3.1), (2) coded issue risks, options, and rationales (§4.3.2), and (3) developed risk, treatment, and rationale taxonomies (§4.3.3).

2. For §4.3.2, coded statements from 10% of the T&S issue discussions were provided to the other annotator, who independently coded the statement. The Kappa coefficients for each type code are: 0.89 for *risk*, 1.0 for *option*, 1.0 for *rationale*, and 1.0 for *chosen*.

3. For §4.3.3: (1) No agreement process was performed for SMP feature list due to the straightforward nature of the list. (2) For the risk statements, a low Kappa score convinced the other annotator and I to independently code all statements and then resolve disagreements. (3) For the pattern taxonomy, a Kappa score of 0.73 was achieved. (4) For the rationale taxonomy, a Kappa score of 0.81 was achieved.

4. Based on the "substantial" agreement between myself and the other annotator, my annotations were used for the remaining 90% of the data except the risk statements.

5. RESULTS

5.1 RQ1: What are the characteristics of T&S issues?

A metadata analysis of T&S issues was performed to study the context in which these issues arise. Specifically I determined when and where they appear over time, which SMP features they present in, and what phase of the SDLC they involve.

5.1.1 Longitudinal Analysis

Figure 5.1 displays the percent of all issues and T&S issues created relative to their respective populations. Both Diaspora and Mastodon saw T&S concerns rise roughly 1-2 years after their respective creation dates with continued persistence over time.

Additional contextual analysis was carried out in each T&S issue:

- Issues were labeled as *feature requests* if they raised concerns with the design or *bugs* if they raised concerns with the implementation. More than 90% were *feature requests* rather than *bugs*.

- It was found that 13 out of 60 T&S issues referenced other SMPs during the issue discussion.[1] Many of these references discuss unwanted behavior that should be avoided. Research has shown that referencing past failures can influence design [81].

5.1.2 Feature Breakdown

Table 5.1 shows the involved platform features and their frequency. The *moderation* and *content sharing* features appeared most frequently, followed by *user registration*. Each feature was also categorized into an element from Smith's honeycomb model (*identity*, *presence*, *relationships*, *reputation*, *groups*, *conversations*, and *sharing*) [20]. Note that the *infrastructure* element was added to account for internal features that users do not interact with.

[1] ↑List of SMPs: Cloudfare, YouTube, Twitter, Facebook, Roblox, Nintendo, Discord, Numerous, Snapchat, Tumblr, Linkedin, BoingBoing, Spotify, Instagram, Kik, Substack, Ravelry, Craigslist, Vimeo, Friendster, Lyft, Anchor.fm, Telegram, GIPHY, Bumble, Gab, Chatroulette, Valve, Twitch, Parler, Dropbox, Yelp, SoundCloud, Livejournal, Convey, Xhamster, Scopely, GoFundMe, Whisper, Wikipedia, Usenet, TikTok, Talkspace, Ask.fm, Pinterest, Amazon, Nextdoor, Disney, Minecraft .

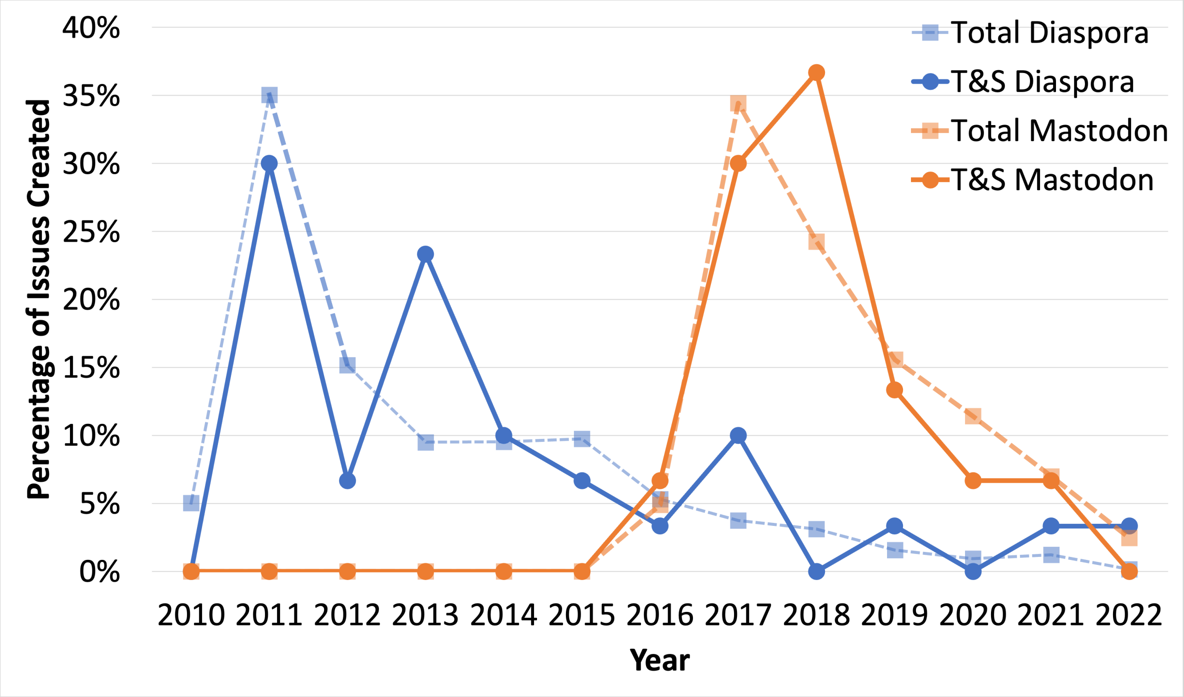

Figure 5.1. Proportion of issues created over time, by SMP (orange–Mastodon; blue–Diaspora) and by type (solid–sampled T&S issues; dashed–all issues). The "all issues" (dashed) and T&S lines (solid) have similar reliability growth curves [82], but the T&S trend seems delayed by 1–2 years. For Diaspora, there are localized spikes in T&S activity several years after the initial peak that ends with an upwards slope in 2021–2022. For Mastodon, the initial peak occurred at age 2 with a resurgence in 2021.

Table 5.1. SMP feature list. Features were openly coded per issue after all T&S issues were gathered. Determinations were based on the primary functionality involved in the issue discussion. Element taxonomy follows Smith [20], with additions in **bold**.

Feature	**Element(s) [20]**	**Description**	**Diaspora**	**Mastodon**	**Total**
Moderation	**Infrastructure**	Moderators: monitor content and enforce platform guidelines	4	8	12
Content sharing	Sharing	Users: post content for others to see	9	2	11
User registration	Identity	Account creation, verification, and on-boarding	6	3	9
Private messaging	Conversations, Groups	Users: Direct communication between two or more users	3	3	6
Content tagging	Sharing	Users: apply labels to their content for discoverability	3	2	5
User relationships	Relationships	Users: follow/friend other users	4	1	5
Content filters	Sharing	Users: hide unwanted content	0	4	4
User filters	Presence, Relationships	Users: prohibit or ignore communication with other users	0	3	3
Instance filters	Groups	Users: prohibit or ignore communication with other instances	0	2	2
Content metadata	Sharing	Users: attach metadata to posted content	0	2	2
User profile	Identity	Users: create a page about themselves	1	0	1

There are some noteworthy differences by platform in each feature's T&S involvement over time. One year after Diaspora's creation, there were a significant number of *content sharing* T&S issues, indicating that this feature posed many T&S risks to the system. In contrast, the early T&S concerns in Mastodon were *moderation*, *content filters*, and *instance filters* issues. The frequency of *user registration* issues remained consistent over time, indicating recurring T&S issues in this feature in both platforms.

5.1.3 Key Findings

Both projects see T&S issue frequency rise 1-2 years after project creation. The *moderation*, *content sharing*, and *user registration* features are most commonly discussed in T&S issues. The *content sharing* and *moderation* features saw respective peaks in activity in 2011 and 2017–2019, respectively. 92% of T&S issues were feature requests instead of bugs. 13 out of 60 T&S issues referenced other SMPs.

5.2 RQ2: What risks are identified in T&S issues?

5.2.1 Threat Actor Analysis

The *threat actor* that each risk statement implicated was analyzed. Among them were *user*, *moderator* (which includes content moderators and server administrators), *bot*, and *external actor*. Over half of risk statements implicated *users* as the primary threat actor. *Moderators* occurred ∼20% of the time, with *bots* and *external actors* comprising the rest. Examples of each threat actor follow:

- *User*: "The captcha will remind the user that this is quite serious and will avoid spamming." (Diaspora #4711)
- *Moderator*: "Moderators [can] access private [content]" (Mastodon #6986)
- *Bot*: "The current one is very bad at preventing bot registrations." (Diaspora #8342)
- *External Actor*: "...risk of a hostile instance harvesting the private messages of unlocked users." (Mastodon #4296)

Table 5.2. T&S risks identified in each repository. Taxonomy adapted from Thomas *et al.* [13] with additions in **bold**.

Risk [13]	**Description**	**Diaspora**	**Mastodon**	**Total**
Toxic Content	Content that users do not wish to see.	5	22	27
Content Leakage	Leak private content to wider audience.	19	5	24
Undermoderation	Moderation that is slow or ineffective.	6	11	17
Overloading	Force target to deal with a sudden influx of content.	6	11	17
Other	Risks that do not fit into any other category.	5	11	16
False reporting	Use of content reporting system with malintent.	6	6	12
Impersonation / Faulty Accounts	Deceive others about identity.	5	5	10
Lockout and Control	Interfere with access to a user's account or any data.	3	3	6
Overmoderation	Moderation that is too invasive or drastic.	2	3	5
Surveillance	Aggregate or monitor user data.	1	2	3

5.2.2 Risk Taxonomy

The *risk identification* step of the ISO risk management process [40] was also carried out. Table 5.2 displays the risk taxonomy and frequencies across 137 risk statements. *Toxic content* is of particular interest in Mastodon, while Diaspora is most concerned with *content leakage*. Mastodon also saw more mentions of *under moderation* concerns rather than *over moderation*. These differences suggest that Mastodon is more focused on unwanted content on the platform and moderation resources to handle T&S risks. Meanwhile, Diaspora values data protection and respecting user privacy.

5.2.3 Risk Statements Over Time

Examining risk statements over time, mentions of *toxic content* peaked from 2016–2019, which contained 24 statements — only 3 were from Diaspora. However, the *content leakage* risk saw an initial spike from Diaspora from 2011–2014, but another wave of activity began in 2016 with a crescendo in 2018 (roughly half of the activity went to each platform).

5.2.4 Risk Landscapes

Figure 5.2 shows the risks identified in the two most common features, *moderation* and *content sharing*. These landscapes are distinct, with *moderation* issues mostly containing *false reporting* and *undermoderation* risk statements and *content sharing* discussions mainly mentioning the *content leakage* risk. Perhaps the most significant result is that *moderation* discussions contained 43 risk statements while *content sharing* features contained 13, resulting in 3 times more risk statements in *moderation* discussions.

5.2.5 Key Findings

Users are the most common threat actors followed by *moderators, external actors, and bots*. *Under moderation* and *overloading* are secondary concerns for both platforms. Mastodon is primarily concerned with *toxic content*, while Diaspora focuses on *content leakage*. *Content leakage* was a concern 14 years after Diaspora was created, but a subsequent spike in activity occurred for both platforms in 2016–2018. Risk landscapes can vary significantly based on which feature the issue deals with.

5.3 RQ3: What options are proposed in T&S issues? How are they selected?

To understand the *risk treatment* process, treatment patterns and their rationales were identified and the effectiveness of the process itself was assessed.

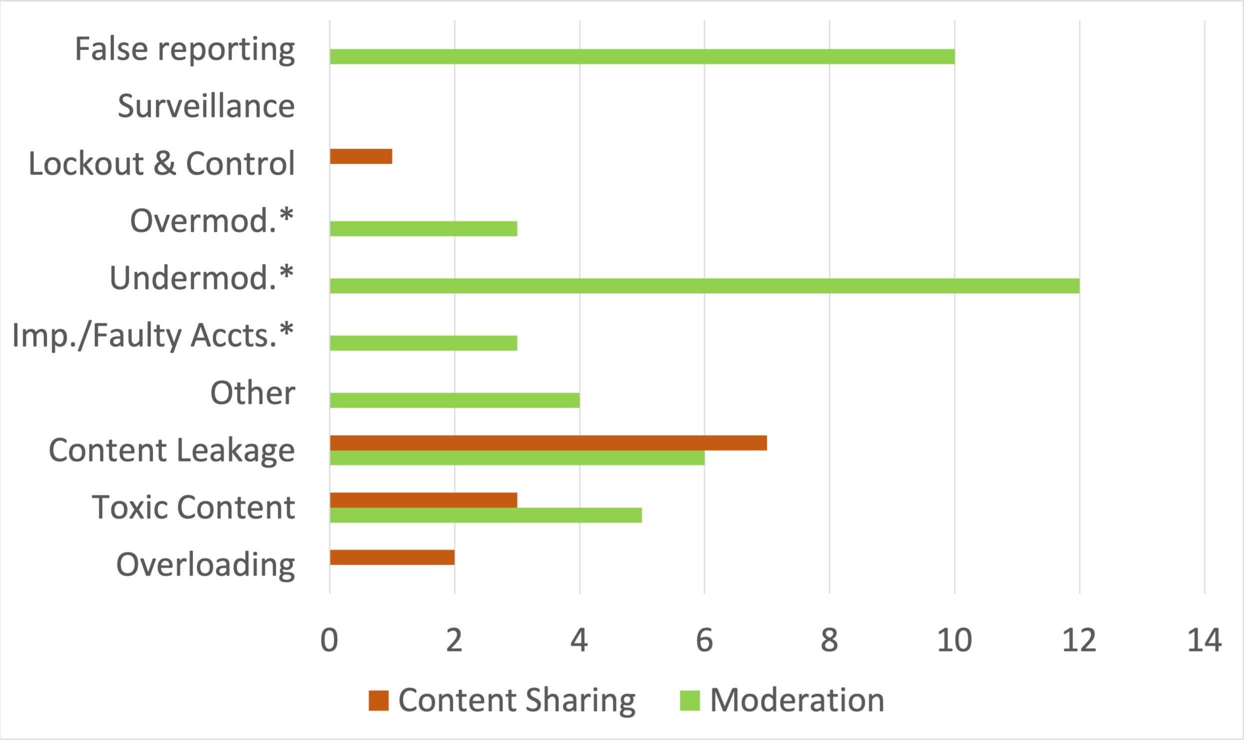

Figure 5.2. Risk landscapes for the *moderation* and *content sharing* features. *Moderation* issues primarily focus on risks of *false reporting* and *undermoderation*. *Content sharing* issues mostly discuss the risk of *content leakage* and there are significantly fewer risks mentioned per issue.

5.3.1 Treatment Taxonomy

To study treatment patterns, thematic coding revealed *options* that treat T&S issues. These overarching themes are termed as *T&S Engineering patterns*. The initial taxonomy was adopted from Koscik [18] and extended in this study (Table 5.3).

First, this study considers the options that were proposed by discussion members. Table 5.3 displays each pattern and the frequencies. *Add moderation* is the most frequently proposed pattern, followed by *require consent*.

Based on the definitions of each pattern, Figure 5.3 superimposes onto the previous context diagram (Figure 2.1) to indicate (a) when each pattern intervenes and (b) whom each pattern relies on. The diagram is split based on *proactive* patterns that intervene before an interaction occurs and *reactive* patterns that intervene afterward. Additionally, the color signifies the party that each pattern relies on. 7 of the identified patterns are *proactive* in nature, while 5 are *reactive*. 4 patterns rely on humans, but the other 8 are fully automated.

By platform, Diaspora sees more *require consent* proposals along with *remove data* and *interaction intervention*. By contrast, Mastodon saw 16 *improve filters* suggestions compared to Diaspora's zero, and more *moderation transparency* proposals. This comparison indicates that Diaspora is more focused on *reactive* patterns, while Mastodon is more concerned with *proactive* ones.

5.3.2 Rationale Taxonomy

Next, the patterns that are actually chosen by engineers and the rationales for those decisions are analyzed. Figure 5.3 shows that *proactive* patterns are chosen less frequently by engineers and chosen options rely on users or moderators more often than not. Table 5.3 indicates that *improve registration* had the highest acceptance rate (38%) and *add moderation* had the lowest (23%). Table 5.4 compares the most common reasons for acceptance (*merged*) or rejection (*no action*) of a proposal.

Among the set of proposed treatments, they tend to be proactive (not reactive) and automated (not relying on humans). However, most chosen options are reactive and do rely on humans. This suggests that engineers select from a preferred minority of solution

Table 5.3. T&S Engineering Patterns. Patterns were coded from *option* statements that treat T&S issues. Parenthesized digits are the total number of occurrences while un-parenthesized are the number of unique issues each pattern appears in. P/R indicates proactive vs. reactive patterns. **Bold** patterns are an addition to the taxonomy adapted from Koscik [18]. See Figure 5.3 for a context diagram with the patterns.

Pattern	Description	Example	P/R	Proposed	Chosen
Add moderation	Add or improve moderation tools	“User groups with ACL would be great though, so we could have multiple admins and/or a moderation team with access to reports." (Mastodon #811)	R	20 (35)	7 (8)
Require consent	Ask for approval from involved stakeholders	“I think it should be unchecked, for privacy reasons." (Diaspora #4343)	P	15 (21)	4 (4)
Improve filters	Allow users to better control the content they see	“I think the exploitation can be reduced arbitrarily to any personal preference by selecting from whom there can be invites." (Mastodon #7369)	R	7 (16)	3 (4)
Reduce visibility	Limit when a feature can be used	“Users should not be allowed to invite users who have blocked or muted them." (Mastodon #7369)	P	8 (12)	1 (1)
Improve registration	Bolster user trustworthiness checks	“About spam, what about a captcha during registration?" (Diaspora #4616)	P	6 (8)	3 (3)
Reduce audience	Limit exposure of content	“Or should their future participations in the conversation be hidden from the view of the person who has ignored them?" (Diaspora #7612)	R	6 (7)	1 (1)
Interaction intervention	Intervene before users contact others	“I think we should add a captcha when reporting a post." (Diaspora #4711)	P	3 (5)	1 (1)
Moderation transparency	Increase clarity of moderation decisions	“...add [moderation] information to the About page... so that people who are vulnerable to harassment can view [it] before deciding... to join." (Mastodon #8557)	R	2 (5)	0 (0)
Interaction transparency	Clarity of events that occurred between users	“In that case (to avoid harrassment) the tagged user should still be notified." (Mastodon #649)	R	3 (4)	0 (0)
Remove data	Remove unnecessary data from platform	“Is it even possible not to generate OpenGraph info, if the post is marked as NSFW?" (Diaspora #7962)	P	4 (4)	0 (0)
Reduce interaction	Limit how a feature can be used	“Blocking someone should make it so that any of their replies to your posts should no longer be considered threaded" (Mastodon #1669)	P	2 (2)	0 (0)
Remove feature	Take out feature	—	P	0 (0)	0 (0)

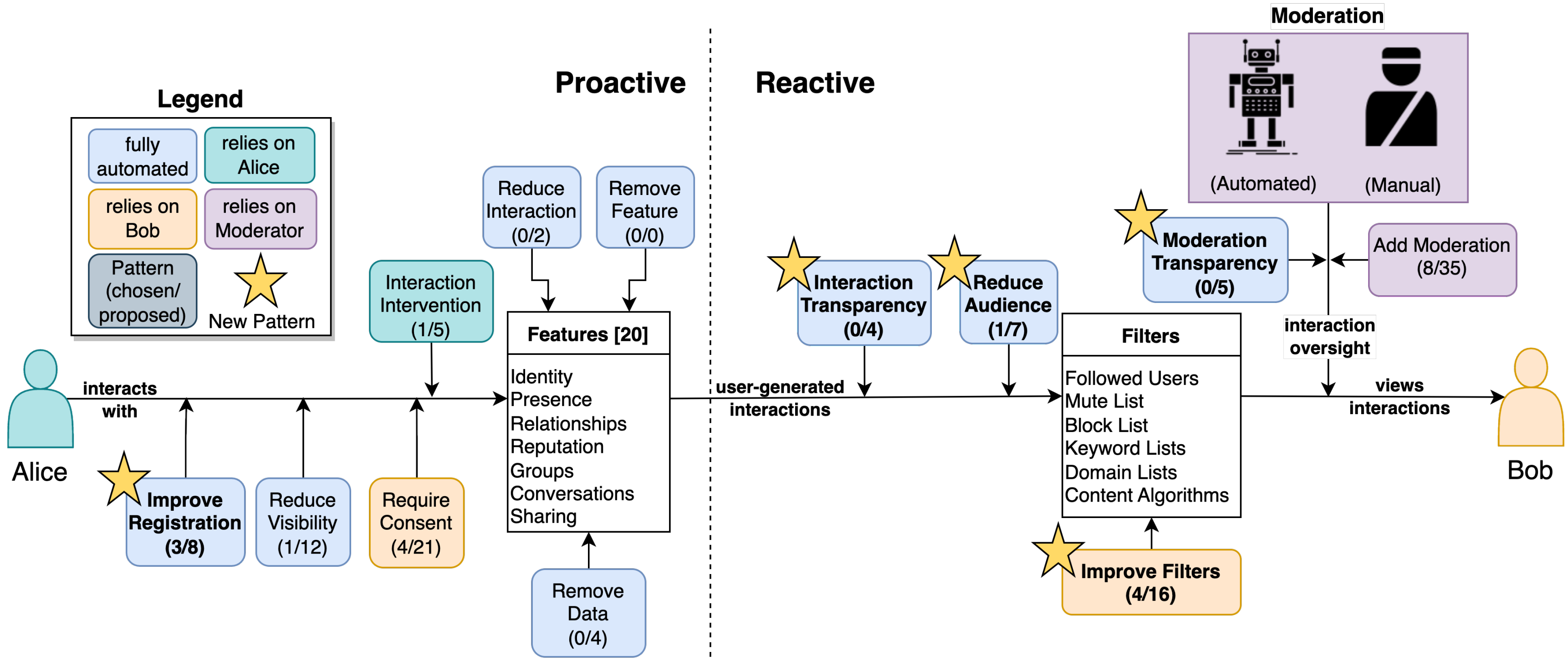

Figure 5.3. SMP context diagram with T&S Engineering patterns. Patterns that intervene before Alice interacts are *proactive* and are otherwise *reactive*. Patterns are fully automated if they do not require user or moderator cooperation to mitigate T&S risk. See Table 5.3 for definitions and frequencies of patterns.

Table 5.4. T&S risk treatment rationales. The general taxonomy from Ko *et al.* is contextualized to T&S [64]. New categories are in **bold**.

Result	Rationale	Description	Count
MERGED	**Safety**	Protects user from T&S risks	14
	Efficiency	Easy completion of tasks	12
	Mod. efficiency	Easy completion of admin/mod tasks	9
	User efficiency	Easy completion of SMP user tasks	3
	Feasibility	Ease of implementation	6
	Flexibility	Handles a variety of use cases	6
	Clarity	Provides clear experience to users	4
	Security	Prevents unwanted data access	3
	Annoyance	Removes hindrance to user activity	2
NO ACTION	**Infeasibility**	Difficulty of implementation	15
	Internal Infeasibility	Difficulty due to internal factors	9
	External Infeasibility	Difficulty due to external factors	6
	Unsafety	Adverse effect to user T&S	9
	Insecure	Susceptible to unwanted data access	5
	Inconsistency	Conflicts with design or user expectations	3
	Uncertainty	Unclear design or T&S environment	1
	Annoyance	Unnecessary hindrance to user activity	1
	Unclarity	Convoluted user experience	1

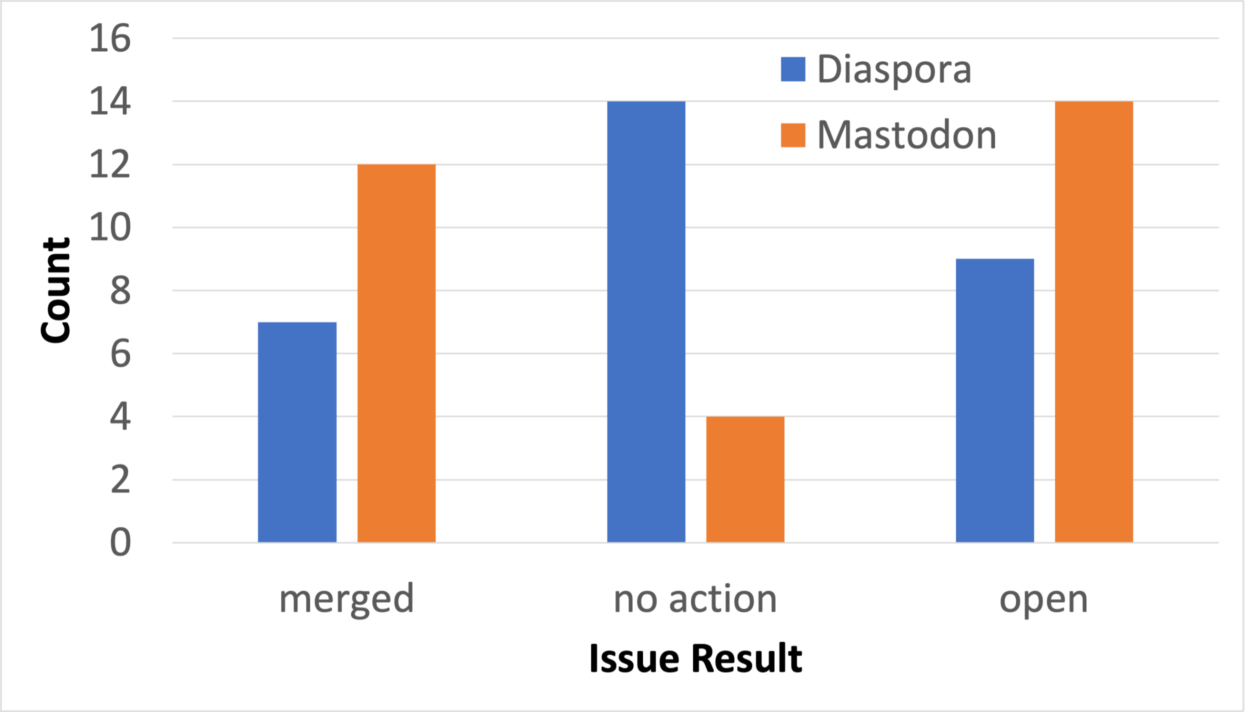

Figure 5.4. Issue result distribution. *Merged* means an issue was closed with some change to the codebase. *No action* means an issue was closed with no change to the codebase. *Open* means the issue is still under discussion.

patterns. Moderator efficiency was cited in many accepted proposals (*e.g.,* supporting human intervention), while federation incompatibility was a common reason to take no action on an opened issue (*e.g.,* preventing automation). The safety and feasiblity of proposals were frequent rationales for both acceptance and rejection.

5.3.3 Issue Resolution Rates

Last, the T&S issue status and age is analyzed to get a sense of how good the T&S risk treatment process is at resolving issues quickly and effectively. It was found that more than one-third of T&S issues are still open with no resolution (Figure 5.4). Closed T&S issues took almost 5 months longer to resolve than the average issue closure time. Figure 5.4 also shows that Diaspora has closed issues with *no action* more frequently than Mastodon.

5.3.4 Key Findings

Add moderation is the most commonly proposed pattern followed by *require consent* and *improve filters* (Table 5.3). The majority of discovered patterns are *proactive* in nature (Table 5.3). Diaspora sees more of these proposals than Mastodon, which had a heavy emphasis on the *improve filters* pattern (Table 5.3). *Reactive* patterns are chosen 13/22 times and those that involve humans are chosen 17/22 times (Table 5.3). Most accepted treatments were associated with *safety* and *moderation efficiency* (Table 5.4). Rejected treatments were commonly *unsafe* for users, *infeasible*, or exhibited *federation incompatibility* (Table 5.4). T&S issues take 147 days longer to resolve compared to the average closure time (Figure 5.4). 38% of identified T&S issues remain open (Figure 5.4). Diaspora is less prone (53%) than Mastodon (87%) to making platform changes in response to T&S issues (Figure 5.4).

6. DISCUSSION

Findings from the study provide larger insights into how OSS SMPs are managing T&S risk. I provide suggestions for how they could be improved, relate findings to a commercial example, and discuss future work to be done.

6.1 Recommendations for OSS SMPs

Several ways in which OSS SMPs might improve their T&S risk management process are discussed.

6.1.1 Communicate Existing Risks

§5.3.4 shows that T&S issues are difficult to resolve. With many of them still open, users are exposed to T&S risks every day, so remaining transparent is critical. Furthermore, transparency-related patterns (*i.e., interaction transparency* and *moderation transparency*) were never chosen by engineers. Evaluating these residual risks, estimating their magnitude, and making users aware of them will reduce their impact. In practice, Meta [83], Twitter [84], and TikTok [85] maintain help centers to provide users with best practices to avoid harmful situations online.

6.1.2 Formalize T&S Reporting

To the best of my knowledge, OSS SMPs are not leveraging data related to T&S on their platforms. While rates of online abuse have been investigated by various third parties such as TSPA [86] and Pew Research [6], first-party data will provide clarity into the current T&S risk landscape. Methods for OSS SMP engineers to collect real T&S data on their platform can provide a clear view of how pervasive T&S issues are and measure results of risk management efforts. Although data is generalized, Meta [87], Snapchat [88], and Discord [89] provide the public with transparency reports. It is reasonable to assume they leverage detailed reporting internally to understand and respond to changing T&S risk environments.

6.1.3 Document Risks and Treatments

In the OSS SMPs that were studied, knowledge of risky features, risk factors, and risk treatments is distributed across project personnel and documents (*e.g.,* distinct issues). The study identified patterns within these features (Table 5.1), factors (Table 5.2), treatments (Table 5.3), and rationales (Table 5.4). Patterns can accelerate the engineering process: prior conversations and decisions could be tracked to guide future T&S discussions. This would promote consistency in decision-making and let precedent resolve dispute.

6.1.4 Explore Proactive Solutions

37 of 60 T&S issues were resolved with a change. As illustrated in Figure 5.3, most of these solution approaches were reactive rather than proactive, despite the majority of them being the latter (Table 5.3). They generally *shared risk* between users and moderators of the system. §6.3.5 discusses how engineers can leverage *T&S By Design* to inform proactive solution development.

6.1.5 Stay Vigilant

The analysis of the T&S defect arrival rate (Figure 5.1) showed that T&S issues manifest later than other defects, and remain present throughout SMP lifespans. As a non-functional requirement similar to cybersecurity, T&S will likely remain a concern for the lifetime of the project and deserve proper attention from engineers. Groups like the Trust & Safety Professional Association [30], the Trust & Safety Foundation [86], and the Trust & Safety Journal [9] exist because promoting T&S is a complex, endless pursuit that requires effort from a variety of stakeholders, including engineers.

6.2 Relations to Commercial SMPs: A Case Study of Young Users on TikTok

While this study of OSS SMPs may not generalize to commercial SMPs, a case study shows potential overlap between the two contexts. I select a critical issue within TikTok, discuss events that have transpired, how TikTok responded, and how these actions relate to my findings.

6.2.1 Background

TikTok has enforced a minimum age for users based on the laws of each country it operates in [90]. However, underage users and exposure to unsafe content has been a consistent problem for the platform. In late 2019, an article from ABC reported that youths as young as 9 years old were using TikTok and were exposed to inappropriate content [91]. In 2020, New York Times reported that "a third of TikTok's U.S. users may be 14 or under" and that many underage users lie about their age when creating an account [92]. In 2021, a study found that 25% of kids 9-17 reported having had a sexually explicit interaction with someone they thought was 18 or older [93]. Later that same year, a 12-year-old died while engaging with a viral TikTok trend called the "Blackout Challenge" where users choke themselves until they pass out [94]. In late 2022, a study reported its results after setting up fake TikTok accounts at the minimum age of 13 – the fake account's feed contained suicide and eating disorder content within minutes of account creation [95]. In April 2023, TikTok was fined £12.7 million by the U.K.'s Information Commissioner's Office for misusing data of young users [96].

How has TikTok responded to these critical societal threats? To answer this question, I consult news releases with the "Safety" tag from the TikTok newsroom page [97] that contain some mention of young or underage TikTok users. TikTok's first news release was about this issue came in 2019 and provided general tips for parents to protect their children including blocking users, leveraging device-level parental controls, encouraging young users to restrict comments, and turn on comment filtering [98]. Later in 2019, TikTok announced the "TikTok for Younger Users" feature that limits sharing, comments, and other interactions [99]. In early 2020, the "Family Pairing" feature was announced that allowed adults to control existing protection features for their child's account including restricting direct messaging, screen time limits, and disabling image and video in direct messages [100]. About a year later in 2021, TikTok summarized its existing work to protect young users including screen limitations, requiring manual birthdate entry, underage account takedowns, and TikTok Live restrictions [101] [1]. Later in 2021, Evans states that TikTok would like to "further enhance proactive protections" and includes pop-ups for young users when posting their first video, disabling posting of public videos, and disabling video downloads for users under 16 [103]. In late 2021, TikTok also posted that new educational resources are made available to parents as part of their "Family Pairing" feature [104]. Finally in March 2023, TikTok announced new features primarily focused on limiting screen time by enforcing a 1 hour time limit for all users under 18 and enforcing push notification schedules for young users.

6.2.2 Analysis

Table 6.1 summarizes TikTok's actions to protect younger users, spanning across the *moderation, content sharing, user registration, private messaging, content filtering, and user filtering* features. Initially, TikTok encouraged parents to use existing features like user blocking and content reporting (see rows with no associated pattern). Soon after, the SMP recognized that customized features were required to address an array of threats. Starting with proactive approaches, the *improve registration* pattern has been used to simply prevent underage users from registering while the *reduce visibility* pattern has limited or disabled commenting, private messaging, and sharing features. Various strategies have *required consent* of teens to control who can interact with their content and parents to control which features their teens can interact with. On the reactive side, TikTok has encouraged users to *improve filtering* of comments and other content they see. Throughout this process, TikTok has also *added moderation* by continually taking down underage accounts and maintained *moderation transparnecy* by publishing routine reports.

TikTok has leveraged several proactive approaches to protect younger users primarily by limiting if, when, and how the *private messaging and content sharing* features can be used. By leveraging the model provided in Figure 5.3, existing strategies can be contextualized so that current actions can be better understood and new solutions can be more effectively

[1]↑ Interestingly, another nearly identical article was published on the same day that changed the wording of some subsections. Most notably, the first article used the phrase "preventing underage people from signing up" [101] while the other uses the phrase "promoting age-appropriate experiences" [102].

Table 6.1. How TikTok has protected young users. Actions span across the *moderation, content sharing, user registration, private messaging, content filtering, and user filtering* features.

Feature	**Year**	**Action**	**Pattern(s)**
Moderation	2019	Report content or a profile directly from within the app	—
	2021	Underage account takedowns	add moderation
	2021	Share information regarding removals of suspected underage accounts	moderation transparency
Content Sharing	2019	Allow comments from followers only	—
	2019	Control who can duet or react to videos	—
	2019	Disable sharing/commenting if under 13	reduce visibility
	2021	Disable TikTok LIVE for young users	reduce visibility
	2021	Make account private	require consent
	2021	Disable Duet and Stitch if under 16	reduce visibility
	2021	Disable downloads on content from accounts under 16	reduce visibility
	2021	Pop-up when teenagers first post to choose visibility level	interaction intervention
User Registration	2019	Enforce age-appropriate experiences	reduce interaction
	2021	Require manual entry of full birthdate	interaction intervention
Private Messaging	2019	Teens can only get messages from followers	reduce visibility
	2019	Parents can disable messaging entirely from privacy settings	require consent, reduce visibility
	2020	Disable images or videos in messages	reduce interaction
	2020	Disable direct messages for accounts under 16	reduce visibility
	2021	Default direct messaging setting to 'no one' for ages 16-17	reduce visibility
Content Filters	2019	Enable comment filters	—
	2021	No notifications after 9pm if age 13-15	
	2021	No notifications after 10pm if age 16-17	reduce visibility
	2023	Forced 60-minute time limit if under 18	reduce visibility
User Filters	2019	Removing unwanted followers	—
	2019	Block unwanted users	—

developed. Many of the mentioned efforts take a perspective where Alice is the young user to protect and Bob is the suspected bad actor that could view their content. Instead, one can change the perspective such that Alice is the suspected bad actor. From this perspective, T&S Engineering patterns take a new tone. In practice this could mean:

- requiring Alice to enable two-factor authentication (improve registration),
- requiring appropriate identification from Alice to view Bob's content (reduce visibility),
- requiring appropriate identification from Alice for Bob to see Alice's content (reduce audience),
- only allowing Alice to react with emojis instead of comment (reduce interaction),
- not listing Alice's account when viewing who has liked a piece of content (reduce interaction), and
- showing Bob when Alice has viewed his video (interaction transparency).

Protecting young users is a critical but challenging effort. While TikTok has taken an array of approaches, a catalog of solution patterns and model can better contextualize current efforts, reveal holes in existing solutions, and standardize the T&S Engineering process.

6.3 Future Work

This exploratory research identified several research opportunities to improve T&S Engineering.

6.3.1 T&S Engineering Pattern Catalog

From §5.2.5 and §5.3.4, there are clear problem and solution themes within SMPs. Properly organizing and tracking these recurrences can prevent some T&S issues from recurring and lead to quicker resolutions when similar problems arise. Tables 5.2 and 5.3 provide the first empirically grounded patterns for T&S problems and solutions in SMPs. Further work in taxonomization, *e.g.,* expanding to more issues or other OSS SMPs (Table 4.1), could improve this catalog. The T&S risks on commercial SMPs could also be incorporated, *e.g.,* following the method of Anandayuvaraj & Davis [105], although the solution patterns are sometimes opaque. Figure 5.3 offers a starting point for organizing such work.

The merits of such a catalog must also be assessed. Context dictates which pattern, if any, may be suitable. I conjecture that T&S risks recur frequently enough within and across SMPs that a pattern catalog would simplify the formulation and selection of T&S risk treatments, resulting in more consistent decisions made more quickly.

6.3.2 Improved T&S Testing

Software testing is a major part of the software development lifecycle. Surprisingly, in the T&S discussions that were studied, *testing was never mentioned.* Operationalizing T&S for automated testing is an open challenge and could be limited [106], [107], however automated procedures could check that basic user boundaries are respected, for example. A possible starting point is the usability testing literature [108], [109]. In addition, tests that require real users can provide a more holistic approach to validating T&S.

6.3.3 Automated Content Moderation in OSS SMPs

Commercial SMPs rely heavily on automated moderation, while OSS SMPs tend to use human moderation. Human moderation has limits. *Under moderation* is a frequent T&S risk in OSS SMPs (Table 5.2). Commercial SMPs have also demonstrated that reliance on human moderation and oversight is not scalable and entails a high cost in the form of inhumane working conditions for their moderators [27], [110]. Both Twitter [23] and YouTube [111] have even shown disinterest in maintaining and prioritizing their T&S teams. However, developing accurate automated moderation has proven challenging because of the amount of contextual information required to make effective judgements. To what extent can automated content moderation be incorporated into OSS SMPs? Is a decentralized OSS SMP instance easier to moderate (*e.g.,* a more homogeneous user base) than a centralized SMP? Investigating automated content moderation could strengthen this weak point in the T&S risk environment. However, there are many T&S considerations to such a proposal, including:

whether and how moderators/users can opt in to this feature; ensuring that data is handled properly; and communicating any other residual risks to involved stakeholders. Furthermore, OSS SMP stakeholders may be unwilling to adopt automated content moderation due to highly-publicized failures in commercial SMPs. Understanding these human factors and the interplay between commercial and OSS SMPs could advance the conversation.

6.3.4 T&S Improvements in Federated Protocols

Federation incompatibility was cited in 7 proposal rejections (Table 5.4) and federation protocols expose OSS SMPs to substantial risk. The largest-scale federation protocol, ActivityPub (8.7M accounts [66]), states the following in their documentation: “While no specific mechanism for combating spam is provided in ActivityPub, it is recommended that servers filter incoming content both by local untrusted users and any remote users through some sort of spam filter” [112]. Adding safety features within the protocol (*e.g.,* anti-spam measures [112]) could increase the feasibility of some T&S treatments on SMPs. End-to-end arguments in system design suggest limits to the T&S impact of a protocol [113], but perhaps some improvement is possible.

6.3.5 T&S By Design

As measured in §5.3.4 and discussed in §6.1.4, many of the observed T&S engineering patterns were reactive, addressing T&S issues by intercepting problematic behavior or content after it has been generated. Designing safe systems may alleviate these issues. However, there is a gap in literature that operationalizes how engineers can do this. This study provides critical steps towards this goal.

The closest work we have to an encompassing T&S design process is the abuse vector mitigation strategies from Koscik [18] and the general principles provided by the *Safety By Design* framework [57] (§2.3.2). This study builds upon the former by providing an empirical basis and adding new design patterns (Table 5.3). Figure 6.1 shows how this study contributes to these prior works. By studying safe software design, practitioners can develop

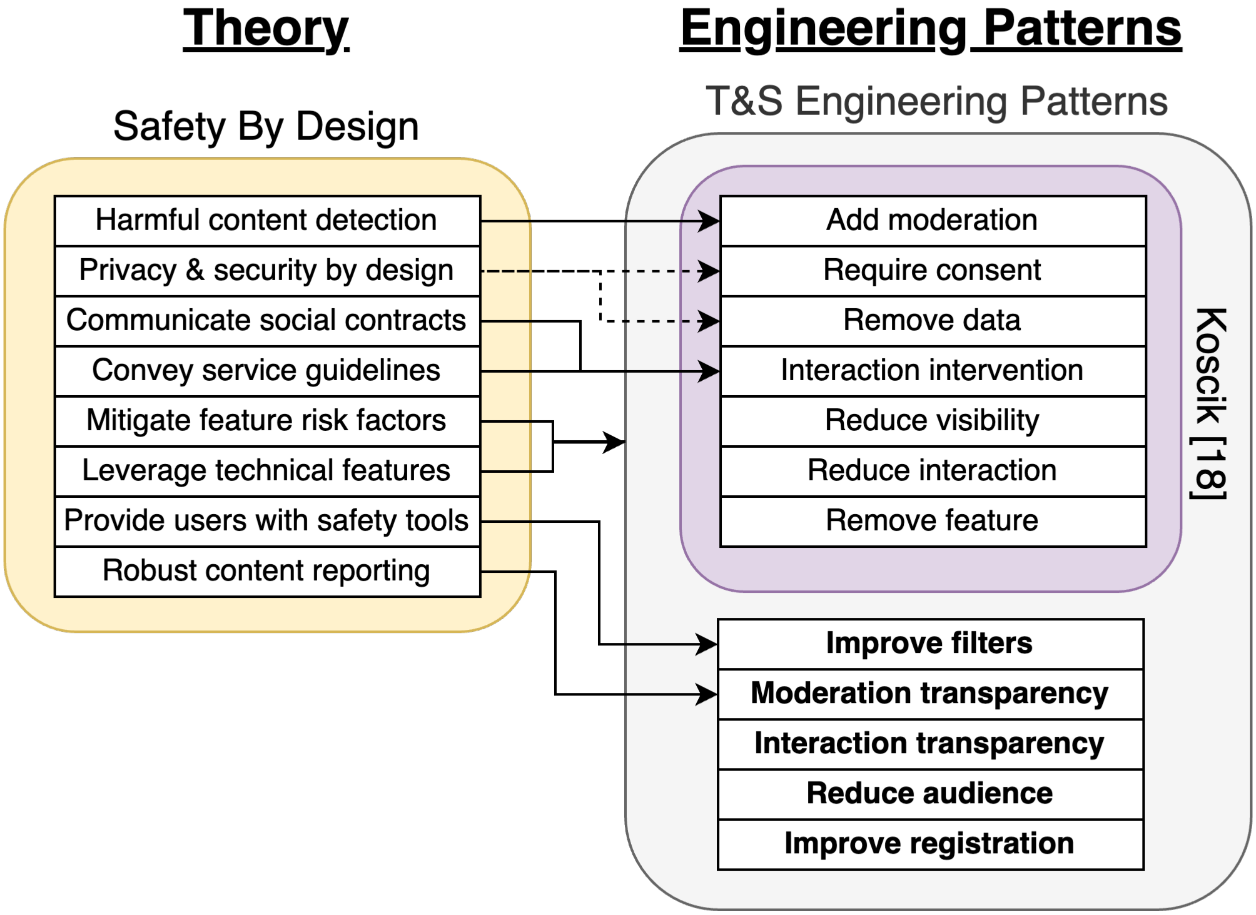

Figure 6.1. Contributions of this study that inform *T&S By Design*. This study builds upon Koscik [18] while reinforcing the Safety By Design framework [57]. Note that the Safety By Design list is not exhaustive and only lists items that involve technical work. Arrows indicate a relation and dashed arrows indicate a partial relation.

measures that are effective [16], [54], scalable, and preventative. The proactive treatment patterns from Table 5.3 provide a starting point for such work.

7. THREATS TO VALIDITY

Internal validity

Methodological choices that could affect my findings. *First*, the work relied on qualitative analysis. To reduce bias, I measured inter-rater agreement with another annotator. To promote comparisons across studies, I used existing taxonomies, extending them as needed. *Second*, the work mined GitHub. This carries concomitant general concerns [114], [115], however there is also a Diaspora-specific concern. The platform uses a separate forum to discuss preliminary feature proposals [73]. Some of these proposals are subsequently filed on GitHub; so I only studied such. Data from this separate forum was omitted because those proposals do not include actions taken by OSS engineers.

External validity

The primary threat to this work is its generalizability. I examined two open-source SMPs with decentralized architectures, omitting other open-source SMPs and all commercial SMPs (which have different goals for their platforms, centralized architectures, and greater resources). I note two mitigating features of the work. First, although the SMPs I studied are a fraction of the size of SMPs such as Facebook, they nevertheless have over 8.5 million accounts — T&S concerns affecting millions of people are worth studying. Second, although open-source, decentralized SMPs were studied, the analysis was built on top of existing taxonomies derived from commercial SMPs. The data fit these taxonomies, suggesting similarities between the contexts. Although in each case, new behaviors were observed and required taxonomy extensions.

As a secondary concern, only N=60 T&S issues were studied. A larger sample size could increase the scope of the findings. I note that 73% of Diaspora issues were analyzed (Table 4.2), indicating that the data was approaching exhaustion for that project. Furthermore, even within this sample, each existing taxonomy was extended, indicating novel findings.

Construct validity

There is no precise definition of “Trust & Safety”. Since T&S is fundamentally a contextual and personal construct, others might reach different conclusions from the data. I operationalized T&S in the terms used by T&S researchers and T&S practitioners (such as TSPA), and used those terms to retrieve relevant issues on GitHub. I then analyzed those issues using my own understanding of T&S risks (chapter 2) by leveraging an ISO risk management standard [40]. However, there is no guarantee that the OSS engineers were using the same terminology. I mitigated this by measuring information retrieval on the keywords.

8. SUMMARY

Social Media Platforms (SMPs) are used by over half the global population. Promoting Trust & Safety (T&S) on SMPs is a major challenge that involves users, moderators, policymakers, and regulators. Software engineering matters too: through design, implementation, and validation, software engineers can reduce an SMP's T&S risks.

I conducted the first empirical study of T&S risks on SMPs from a software engineering perspective. I studied 60 T&S-related GitHub Issues for the two most popular open-source SMPs, Mastodon and Diaspora. My work identified novel SMP risks, engineering patterns, and resolution rationales. The key findings are: (1) T&S issues persist throughout a platform's lifetime and mostly require design changes; (2) T&S issues are hard to resolve or remain open; (3) Selected treatments are mostly reactive, waiting until T&S risks manifest to intervene; and (4) Selected treatments mostly share risk with users or moderators, despite many alternatives. This work suggests that, in open-source SMPs, there is currently no systematic engineering approach to promoting T&S. I show opportunities for research on software design, decision-making, and validation for T&S in SMPs.

REFERENCES

[1] DataReportal, *Global Social Media Statistics*, Jan. 2023. [Online]. Available: https://datareportal.com/social-media-users (visited on 03/02/2023).

[2] A. Whiting and D. Williams, "Why people use social media: A uses and gratifications approach," *Qualitative Market Research: An International Journal*, vol. 16, no. 4, pp. 362–369, Jan. 2013, ISSN: 1352-2752. DOI: 10.1108/QMR-06-2013-0041. (visited on 04/20/2022).

[3] S. E. Rolland and G. Parmentier, "The Benefit of Social Media: Bulletin Board Focus Groups as a Tool for Co-creation," *International Journal of Market Research*, vol. 55, no. 6, pp. 809–827, Nov. 2013, ISSN: 1470-7853. DOI: 10.2501/IJMR-2013-068. (visited on 04/18/2022).

[4] R. Deloatch, B. P. Bailey, A. Kirlik, and C. Zilles, "I Need Your Encouragement! Requesting Supportive Comments on Social Media Reduces Test Anxiety," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, ser. CHI '17, New York, NY, USA: Association for Computing Machinery, May 2017, pp. 736–747, ISBN: 978-1-4503-4655-9. DOI: 10.1145/3025453.3025709. (visited on 04/18/2022).

[5] Z. Ashktorab and J. Vitak, "Designing Cyberbullying Mitigation and Prevention Solutions through Participatory Design With Teenagers," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, ser. CHI '16, New York, NY, USA: Association for Computing Machinery, May 2016, pp. 3895–3905, ISBN: 978-1-4503-3362-7. DOI: 10.1145/2858036.2858548. (visited on 04/18/2022).

[6] E. A. Vogels, "The State of Online Harassment," Pew Research Center, Tech. Rep., Jan. 2021. (visited on 11/02/2022).

[7] A. Marwick, B. Clancy, and K. Furl, "Far-Right Online Radicalization: A Review of the Literature," *The Bulletin of Technology & Public Life*, May 2022. DOI: 10.21428/bfcb0bff.e9492a11. (visited on 11/04/2022).

[8] M. Singhal, C. Ling, P. Paudel, *et al.*, *SoK: Content Moderation in Social Media, from Guidelines to Enforcement, and Research to Practice*, Oct. 2022. DOI: 10.48550/arXiv.2206.14855. arXiv: 2206.14855 [cs]. (visited on 11/02/2022).

[9] E. Cryst, S. Grossman, J. Hancock, A. Stamos, and D. Thiel, “Introducing the Journal of Online Trust and Safety,” *Journal of Online Trust and Safety*, vol. 1, no. 1, Oct. 2021, ISSN: 2770-3142. (visited on 07/18/2022).

[10] L. Galantino, *Trust & Safety Engineering @ GitHub*, May 2019. (visited on 06/11/2022).

[11] M. Fire, R. Goldschmidt, and Y. Elovici, “Online Social Networks: Threats and Solutions,” *IEEE Communications Surveys Tutorials*, vol. 16, no. 4, pp. 2019–2036, 2014, ISSN: 1553-877X. DOI: 10.1109/COMST.2014.2321628.

[12] A. M. Memon, S. G. Sharma, S. S. Mohite, and S. Jain, “The role of online social networking on deliberate self-harm and suicidality in adolescents: A systematized review of literature,” *Indian Journal of Psychiatry*, vol. 60, no. 4, pp. 384–392, 2018, ISSN: 0019-5545. DOI: 10.4103/psychiatry.IndianJPsychiatry_414_17. (visited on 11/02/2022).

[13] K. Thomas, D. Akhawe, M. Bailey, *et al.*, “SoK: Hate, Harassment, and the Changing Landscape of Online Abuse,” in *2021 IEEE Symposium on Security and Privacy (SP)*, May 2021, pp. 247–267. DOI: 10.1109/SP40001.2021.00028.

[14] J. Qian, A. Bethke, Y. Liu, E. Belding, and W. Y. Wang, “A Benchmark Dataset for Learning to Intervene in Online Hate Speech,” *arXiv:1909.04251 [cs]*, Sep. 2019. arXiv: 1909.04251 [cs]. (visited on 04/01/2021).

[15] D. Kiela, H. Firooz, A. Mohan, *et al.*, “The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes,” *arXiv:2005.04790 [cs]*, Jun. 2020. arXiv: 2005.04790 [cs]. (visited on 04/01/2021).

[16] J. Kim, C. McDonald, P. Meosky, M. Katsaros, and T. Tyler, “Promoting Online Civility Through Platform Architecture,” *Journal of Online Trust and Safety*, vol. 1, no. 4, Sep. 2022, ISSN: 2770-3142. DOI: 10.54501/jots.v1i4.54. (visited on 11/22/2022).

[17] S. Arumugam and V. Venugopal, “Detection and Verification of Cloned Profiles in Online Social Networks Using MapReduce Based Clustering and Classification,” *International Journal of Intelligent Systems and Applications in Engineering*, vol. 11, no. 1, pp. 195–207, Jan. 2023, ISSN: 2147-6799. (visited on 01/25/2023).

[18] T. Koscik, *Identifying Abuse Vectors*, 2018. [Online]. Available: https://web.archive.org/web/20220818200307/https://spinecone.gitbooks.io/identifying-abuse-vectors/content/ (visited on 06/02/2022).

[19] C. T. Carr and R. A. Hayes, "Social Media: Defining, Developing, and Divining," *Atlantic Journal of Communication*, vol. 23, no. 1, pp. 46–65, Jan. 2015, ISSN: 1545-6870. DOI: 10.1080/15456870.2015.972282. (visited on 03/08/2022).

[20] G. Smith, *Social Software Building Blocks*, Apr. 2007. [Online]. Available: https://web.archive.org/web/20171123070545/http://nform.com/ideas/social-software-building-blocks (visited on 02/15/2022).

[21] J. H. Kietzmann, K. Hermkens, I. P. McCarthy, and B. S. Silvestre, "Social media? Get serious! Understanding the functional building blocks of social media," *Business Horizons*, SPECIAL ISSUE: SOCIAL MEDIA, vol. 54, no. 3, pp. 241–251, May 2011, ISSN: 0007-6813. DOI: 10.1016/j.bushor.2011.01.005. (visited on 02/27/2022).

[22] *Alexa top 1 million websites*, 2022. [Online]. Available: https://www.expireddomains.net/alexa-top-websites.

[23] M. O'Brien and B. Ortutay, *Musk's Twitter disbands its Trust and Safety advisory group*, Dec. 2022. [Online]. Available: https://apnews.com/article/elon-musk-twitter-inc-technology-business-a9b795e8050de12319b82b5dd7118cd7 (visited on 02/21/2023).

[24] Wikipedia, "Comparison of microblogging and similar services," *Wikipedia*, Apr. 2022. (visited on 10/26/2022).

[25] Wikipedia, "Activity stream," *Wikipedia*, Sep. 2022. (visited on 10/26/2022).

[26] A. Mansoux and R. Roscam Abbing, "Seven theses on the fediverse and the becoming of floss," 2020.

[27] T. S. Mullaney, B. Peters, M. Hicks, and K. Philip, "Your Computer Is on Fire," in *Your Computer Is on Fire*, MIT Press, 2021, pp. 20–21.

[28] eBay, *About eBay: Press Releases*, Jan. 1999. [Online]. Available: https://web.archive.org/web/20000815064513/http://pages.ebay.com/community/aboutebay/releases/9901.html (visited on 03/06/2023).

[29] N. Confessore, "Cambridge Analytica and Facebook: The Scandal and the Fallout So Far," *The New York Times*, Apr. 2018, ISSN: 0362-4331. (visited on 02/02/2023).

[30] Trust and Safety Professional Association, *What We Do*, n.d. [Online]. Available: https://www.tspa.org/what-we-do/ (visited on 02/02/2023).

[31] A. Marwick, *Trust & safety: The formalization of profession*, virtual, Mar. 2021. [Online]. Available: https://tiara.org/wp-content/uploads/2021/07/amarwick_cv072021.pdf.

[32] A. Cai, A. Macgillivray, C. Tsao, D. Dixon, and E. Goldman, *New organizations dedicated to online trust and safety*, https://www.tspa.org/2020/06/17/new-organizations-dedicated-to-online-trust-and-safety, Jun. 2020.

[33] D. Leong, *Adding Community & Safety checks to new features*, Jan. 2017. (visited on 06/11/2022).

[34] Trust and Safety Professional Association, *Senior Security Engineer, Trust & Safety*, Jul. 2022. [Online]. Available: https://web.archive.org/web/20220808140939/https://www.tspa.org/job/senior-security-engineer-trust-safety/ (visited on 11/04/2022).

[35] Cloudflare, *Trust & Safety Engineering Team*, Dec. 2021. [Online]. Available: https://web.archive.org/web/20220128033140/https://www.builtinaustin.com/job/engineer/software-engineer-trust-safety-engineering-team/76783 (visited on 06/11/2022).

[36] M. Samuelson, *How Pinterest built its Trust & Safety team*, Apr. 2022. (visited on 09/21/2022).

[37] S. Xie, *Building a Label-Based Enforcement Pipeline for Trust & Safety*, May 2021. (visited on 09/15/2022).

[38] Trust and Safety Professional Association, *TSPA Job Board*, n.d. [Online]. Available: http://web.archive.org/web/20221219211619/https://www.tspa.org/explore/job-board/ (visited on 11/04/2022).

[39] T. W. House, *Launching the global partnership for action on gender-based online harassment and abuse*, Mar. 2022. [Online]. Available: https://www.whitehouse.gov/gpc/briefing-room/2022/03/18/launching-the-global-partnership-for-action-on-gender-based-online-harassment-and-abuse/.

[40] International Standards Organization, *ISO 31000:2018(en), Risk management Guidelines*, 2018. [Online]. Available: https://www.iso.org/obp/ui/%5C#iso:std:iso:31000 (visited on 07/07/2022).

[41] H. Kumar, S. Jain, and R. Srivastava, "Risk analysis of online social networks," in *2016 International Conference on Computing, Communication and Automation (ICCCA)*, Apr. 2016, pp. 846–851. DOI: 10.1109/CCAA.2016.7813833.

[42] B. Tiganoaia, A. Cernian, and A. Niculescu, "The risks in the social networks An exploratory study," in *2017 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, vol. 2, Sep. 2017, pp. 974–977. DOI: 10.1109/IDAACS.2017.8095232.

[43] G. F. AlMudahi, L. K. AlSwayeh, S. A. AlAnsary, and R. Latif, "Social Media Privacy Issues, Threats, and Risks," in *2022 Fifth International Conference of Women in Data Science at Prince Sultan University (WiDS PSU)*, Mar. 2022, pp. 155–159. DOI: 10.1109/WiDS-PSU54548.2022.00043.

[44] A. A. Hasib, "Threats of online social networks," *International Journal of Computer Science and Network Security*, vol. 9, no. 11, pp. 288–293, 2009.

[45] C. Laorden, B. Sanz, G. Alvarez, and P. G. Bringas, "A Threat Model Approach to Threats and Vulnerabilities in On-line Social Networks," in *Computational Intelligence in Security for Information Systems 2010*, Á. Herrero, E. Corchado, C. Redondo, and Á. Alonso, Eds., ser. Advances in Intelligent and Soft Computing, Berlin, Heidelberg: Springer, 2010.

[46] Y. Wang and R. K. Nepali, "Privacy threat modeling framework for online social networks," in *2015 International Conference on Collaboration Technologies and Systems (CTS)*, Jun. 2015, pp. 358–363. DOI: 10.1109/CTS.2015.7210449.

[47] Y. Wang, M. McKee, A. Torbica, and D. Stuckler, "Systematic Literature Review on the Spread of Health-related Misinformation on Social Media," *Social Science & Medicine*, vol. 240, p. 112 552, Nov. 2019, ISSN: 0277-9536. DOI: 10.1016/j.socscimed.2019.112552. (visited on 04/22/2022).

[48] S. Trabelsi and H. Bouafif, "Abusing social networks with abuse reports: A coalition attack for social networks," in *2013 International Conference on Security and Cryptography (SECRYPT)*, Jul. 2013, pp. 1–6.

[49] W. A. Usmani, D. Marques, I. Beschastnikh, K. Beznosov, T. Guerreiro, and L. Carriço, "Characterizing Social Insider Attacks on Facebook," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, ser. CHI '17, New York, NY, USA: Association for Computing Machinery, May 2017, pp. 3810–3820, ISBN: 978-1-4503-4655-9. DOI: 10.1145/3025453.3025901. (visited on 04/18/2022).

[50] J. M. Such, J. Porter, S. Preibusch, and A. Joinson, “Photo Privacy Conflicts in Social Media: A Large-scale Empirical Study,” in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, ser. CHI ’17, New York, NY, USA: Association for Computing Machinery, May 2017, pp. 3821–3832, ISBN: 978-1-4503-4655-9. DOI: 10.1145/3025453.3025668. (visited on 04/18/2022).

[51] J. Cheng, M. Bernstein, C. Danescu-Niculescu-Mizil, and J. Leskovec, “Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions,” in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, ser. CSCW ’17, New York, NY, USA: Association for Computing Machinery, Feb. 2017, pp. 1217–1230, ISBN: 978-1-4503-4335-0. DOI: 10.1145/2998181.2998213. (visited on 06/02/2022).

[52] N. Kashmar, M. Adda, H. Ibrahim, and M. Atieh, “Access Control in Cybersecurity and Social Media,” *Access Control in Cybersecurity and Social Media*, Feb. 2021.

[53] G. Misra and J. M. Such, “How Socially Aware Are Social Media Privacy Controls?” *Computer*, vol. 49, no. 3, pp. 96–99, Mar. 2016, ISSN: 1558-0814. DOI: 10.1109/MC.2016.83.

[54] J. Seering, T. Fang, L. Damasco, M. ’. Chen, L. Sun, and G. Kaufman, “Designing User Interface Elements to Improve the Quality and Civility of Discourse in Online Commenting Behaviors,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI ’19, New York, NY, USA: Association for Computing Machinery, May 2019, pp. 1–14, ISBN: 978-1-4503-5970-2. DOI: 10.1145/3290605.3300836. (visited on 02/01/2023).

[55] A. Cavoukian and M. Dixon, *Privacy and Security by Design: An Enterprise Architecture Approach*. Information and Privacy Commissioner of Ontario, Canada, 2013.

[56] I. Rubinstein and N. Good, “Privacy by Design: A Counterfactual Analysis of Google and Facebook Privacy Incidents,” *SSRN Electronic Journal*, 2012, ISSN: 1556-5068. DOI: 10.2139/ssrn.2128146. (visited on 01/31/2023).

[57] eSafety Commissioner, *Safety by Design*, n.d. [Online]. Available: https://web.archive.org/web/20220308081249/https://www.esafety.gov.au/industry/safety-by-design (visited on 03/08/2023).

[58] I. Kayes and A. Iamnitchi, “Privacy and security in online social networks: A survey,” *Online Social Networks and Media*, vol. 3–4, pp. 1–21, Oct. 2017, ISSN: 2468-6964. DOI: 10.1016/j.osnem.2017.09.001. (visited on 04/28/2022).

[59] C. Bartolini, A. Calabró, and E. Marchetti, "GDPR and business processes: An effective solution," in *Proceedings of the 2nd International Conference on Applications of Intelligent Systems - APPIS '19*, Las Palmas de Gran Canaria, Spain: ACM Press, 2019, pp. 1–5, ISBN: 978-1-4503-6085-2. DOI: 10.1145/3309772.3309779. (visited on 03/19/2021).

[60] Trust and Safety Professional Association, *Policy Development*, n.d. [Online]. Available: https://www.tspa.org/curriculum/ts-fundamentals/policy/policy-development/ (visited on 02/02/2023).

[61] W. Wang, D. Arya, N. Novielli, J. Cheng, and J. L. Guo, "ArguLens: Anatomy of Community Opinions On Usability Issues Using Argumentation Models," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA: Association for Computing Machinery, Apr. 2020, pp. 1–14, ISBN: 978-1-4503-6708-0. (visited on 05/25/2022).

[62] A. S. Mashiyat, M. Famelis, R. Salay, and M. Chechik, "Using developer conversations to resolve uncertainty in software development: A position paper," in *Proceedings of the 4th International Workshop on Recommendation Systems for Software Engineering*, ser. RSSE 2014, New York, NY, USA: Association for Computing Machinery, Jun. 2014, pp. 1–5, ISBN: 978-1-4503-2845-6. DOI: 10.1145/2593822.2593823. (visited on 06/09/2022).

[63] K. A. Safwan and F. Servant, "Decomposing the rationale of code commits: The software developer's perspective," in *Proceedings of the ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/SIGSOFT FSE 2019, Tallinn, Estonia, August 26-30, 2019*, M. Dumas, D. Pfahl, S. Apel, and A. Russo, Eds., ACM, 2019, pp. 397–408. DOI: 10.1145/3338906.3338979. [Online]. Available: https://doi.org/10.1145/3338906.3338979.

[64] A. J. Ko and P. K. Chilana, "Design, discussion, and dissent in open bug reports," in *Proceedings of the 2011 iConference*, ser. iConference '11, New York, NY, USA: Association for Computing Machinery, Feb. 2011, pp. 106–113, ISBN: 978-1-4503-0121-3. DOI: 10.1145/1940761.1940776. (visited on 07/05/2022).

[65] P. Ralph, N. bin Ali, S. Baltes, *et al.*, "Empirical standards for software engineering research," *arXiv preprint arXiv:2010.03525*, 2020. arXiv: 2010.03525.

[66] The Federation, *The Federation - a statistics hub*, n.d. [Online]. Available: https://the-federation.info/ (visited on 10/26/2022).

[67] K. Huang, "What Is Mastodon and Why Are People Leaving Twitter for It?" *The New York Times*, Nov. 2022, ISSN: 0362-4331. (visited on 03/10/2023).

[68] Q. Xu, J. C. Davis, Y. C. Hu, and A. Jindal, "An Empirical Study on the Impact of Deep Parameters on Mobile App Energy Usage," in *IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*, 2022, p. 12.

[69] W. Jiang, N. Synovic, M. Hyatt, *et al.*, "An Empirical Study of Pre-Trained Model Reuse in the Hugging Face Deep Learning Model Registry," arXiv, Mar. 2023. DOI: 10.48550/arXiv.2303.02552. arXiv: 2303.02552 [cs]. (visited on 03/09/2023).

[70] H. Borges and M. Tulio Valente, "What's in a GitHub Star? Understanding Repository Starring Practices in a Social Coding Platform," *Journal of Systems and Software*, vol. 146, pp. 112–129, Dec. 2018, ISSN: 0164-1212. DOI: 10.1016/j.jss.2018.09.016. (visited on 03/09/2023).

[71] J. Garcia, Y. Feng, J. Shen, S. Almanee, Y. Xia, and Q. A. Chen, "A Comprehensive Study of Autonomous Vehicle Bugs," in *2020 IEEE/ACM 42nd International Conference on Software Engineering (ICSE)*, Oct. 2020, pp. 385–396.

[72] mastodon, *Mastodon/mastodon*, n.d. [Online]. Available: Mastodon (visited on 03/26/2022).

[73] diaspora, *Diaspora/diaspora*, n.d. [Online]. Available: diaspora (visited on 03/26/2022).

[74] Y. Zhang, Y. Chen, S.-C. Cheung, Y. Xiong, and L. Zhang, "An empirical study on TensorFlow program bugs," in *Proceedings of the 27th ACM SIGSOFT International Symposium on Software Testing and Analysis*, Amsterdam Netherlands: ACM, Jul. 2018, pp. 129–140, ISBN: 978-1-4503-5699-2. DOI: 10.1145/3213846.3213866. (visited on 10/04/2021).

[75] A. Makhshari and A. Mesbah, "IoT Bugs and Development Challenges," in *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*, May 2021, pp. 460–472. DOI: 10.1109/ICSE43902.2021.00051.

[76] J. Hancock, *Journal of Online Trust and Safety*, 2022. [Online]. Available: https://tsjournal.org/index.php/jots (visited on 06/07/2022).

[77] D. Arya, W. Wang, J. L. C. Guo, and J. Cheng, “Analysis and detection of information types of open source software issue discussions,” in *Proceedings of the 41st International Conference on Software Engineering*, ser. ICSE ’19, Montreal, Quebec, Canada: IEEE Press, May 2019, pp. 454–464. DOI: 10.1109/ICSE.2019.00058. (visited on 05/25/2022).

[78] G. Viviani, C. Janik-Jones, M. Famelis, and G. Murphy, “The Structure of Software Design Discussions,” in *2018 IEEE/ACM 11th International Workshop on Cooperative and Human Aspects of Software Engineering (CHASE)*, May 2018, pp. 104–107.

[79] International Organization for Standardization, *ISO 26262-1:2018*, Dec. 2018. [Online]. Available: https://www.iso.org/standard/68383.html (visited on 05/04/2022).

[80] M. L. McHugh, “Interrater reliability: The kappa statistic,” *Biochemia Medica*, vol. 22, no. 3, pp. 276–282, Oct. 2012, ISSN: 1330-0962. (visited on 01/10/2023).

[81] D. Anandayuvaraj, P. Thulluri, J. Figueroa, H. Shandilya, and J. C. Davis, *Towards a failure-aware SDLC for internet of things*, 2022.

[82] C. Stringfellow and A. A. Andrews, “An empirical method for selecting software reliability growth models,” *Empirical Software Engineering (EMSE)*, vol. 7, pp. 319–343, 2002.

[83] Meta, *Staying Safe*, n.d. [Online]. Available: https://web.archive.org/web/20220308081255/https://www.facebook.com/help/592679377575472/?helpref=uf%5C_share (visited on 03/08/2023).

[84] Twitter, *Safety and security*, n.d. [Online]. Available: https://web.archive.org/web/20220308081252/https://help.twitter.com/en/safety-and-security (visited on 03/08/2023).

[85] TikTok, *Safety Center*, n.d. [Online]. Available: https://www.tiktok.com/safety/en/ (visited on 03/08/2023).

[86] Trust and Safety Foundation Project, *Case Studies*, n.d. [Online]. Available: http://web.archive.org/web/20230101182825/https://trustandsafetyfoundation.org/case-studies/ (visited on 06/07/2022).

[87] Meta, *Transparency reports / Transparency Center*, n.d. [Online]. Available: https://web.archive.org/web/20220308081251/https://transparency.fb.com/data/ (visited on 03/08/2023).

[88] Snapchat, *Snapchat Transparency Report | Snapchat Transparency*, n.d. [Online]. Available: https://web.archive.org/web/20220308081253/https://values.snap.com/privacy/transparency (visited on 03/08/2023).

[89] Discord, *Transparency Reports*, n.d. [Online]. Available: https://web.archive.org/web/20220308081258/https://discord.com/tags/transparency-reports (visited on 03/08/2023).

[90] TikTok, *Guardian's Guide*, Mar. 2021. [Online]. Available: https://web.archive.org/web/20230413213943/https://www.tiktok.com/safety/en/guardians-guide/ (visited on 04/14/2023).

[91] A. B. C. News, *Young kids could be seeing mature content on TikTok. Here's how to keep them safe*, Oct. 2019. [Online]. Available: https://web.archive.org/web/20220720081827/https://abcnews.go.com/GMA/Living/young-kids-mature-content-tiktok-heres-safe/story?id=66366182 (visited on 04/14/2023).

[92] R. Zhong and S. Frenkel, *A Third of TikTok's U.S. Users May Be 14 or Under, Raising Safety Questions*, Aug. 2020. [Online]. Available: https://web.archive.org/web/20230330091416/https://www.nytimes.com/2020/08/14/technology/tiktok-underage-users-ftc.html (visited on 04/14/2023).

[93] C. Newton, *The child safety problem on platforms is worse than we knew*, May 2021. [Online]. Available: https://web.archive.org/web/20230330144813/https://www.theverge.com/2021/5/12/22432863/child-safety-platforms-thorn-report-snap-facebook-youtube-tiktok (visited on 04/14/2023).

[94] T. Dowd, *This Dangerous TikTok Challenge Just Killed a 12-Year-Old*, Jul. 2021. [Online]. Available: https://web.archive.org/web/20230124210038/https://www.vice.com/en/article/pkbxm9/tiktok-blackout-challenge-kill-children (visited on 04/14/2023).

[95] S. M. Kelly, *TikTok may push potentially harmful content to teens within minutes, study finds | CNN Business*, Dec. 2022. [Online]. Available: https://web.archive.org/web/20230411112722/https://www.cnn.com/2022/12/15/tech/tiktok-teens-study-trnd/index.html (visited on 04/14/2023).

[96] Information Commissioner's Office, *ICO fines TikTok č12.7 million for misusing children's data*, Apr. 2023. [Online]. Available: https://ico.org.uk/about-the-ico/media-centre/news-and-blogs/2023/04/ico-fines-tiktok-127-million-for-misusing-children-s-data/ (visited on 04/16/2023).

[97] TikTok, *TikTok News and Top Stories | TikTok Newsroom*, Aug. 2018. [Online]. Available: https://web.archive.org/web/20230414084706/https://newsroom.tiktok.com/en-us/ (visited on 04/14/2023).

[98] TikTok, *TikTok's Top 10 Tips for Parents*, Oct. 2019. [Online]. Available: https://web.archive.org/web/20221006001304/https://newsroom.tiktok.com/en-us/tiktoks-top-10-tips-for-parents (visited on 04/14/2023).

[99] TikTok, *TikTok for Younger Users*, Dec. 2019. [Online]. Available: https://web.archive.org/web/20230322110612/https://newsroom.tiktok.com/en-us/tiktok-for-younger-users (visited on 04/14/2023).

[100] J. Collins, *TikTok introduces Family Pairing*, Apr. 2020. [Online]. Available: https://web.archive.org/web/20230412022455/https://newsroom.tiktok.com/en-us/tiktok-introduces-family-pairing (visited on 04/14/2023).

[101] T. Elizabeth, *Our work to keep TikTok a place for people 13 and over*, May 2021. [Online]. Available: https://web.archive.org/web/20230315224514/https://newsroom.tiktok.com/en-eu/our-work-to-keep-tiktok-a-place-for-people-13-and-over-eu (visited on 04/14/2023).

[102] T. Elizabeth, *Our work to design an age-appropriate experience on TikTok*, May 2021. [Online]. Available: https://web.archive.org/web/20230130014809/https://newsroom.tiktok.com/en-us/our-work-to-design-an-age-appropriate-experience-on-tiktok/ (visited on 04/14/2023).

[103] A. Evans, *Furthering our safety and privacy commitments for teens on TikTok*, Aug. 2021. [Online]. Available: https://web.archive.org/web/20221018153406/https://newsroom.tiktok.com/en-au/furthering-our-safety-and-privacy-commitments-for-teens-tiktok (visited on 04/14/2023).

[104] A. Evans, *New Family Pairing resources offer digital safety advice from teens*, Sep. 2021. [Online]. Available: https://web.archive.org/web/20230311000442/https://newsroom.tiktok.com/en-us/new-family-pairing-resources-offer-digital-safety-advice-from-teens (visited on 04/14/2023).

[105] D. Anandayuvaraj and J. C. Davis, “Reflecting on recurring failures in iot development,” in *37th IEEE/ACM International Conference on Automated Software Engineering–New Ideas and Emerging Results Track (ASE-NIER'22)*, 2022, pp. 1–5.

[106] A. R. Ibrahimzada, Y. Varli, D. Tekinoglu, and R. Jabbarvand, "Perfect is the enemy of test oracle," in *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/FSE 2022, Singapore, Singapore, November 14-18, 2022*, A. Roychoudhury, C. Cadar, and M. Kim, Eds., ACM, 2022, pp. 70–81. DOI: 10.1145/3540250.3549086. [Online]. Available: https://doi.org/10.1145/3540250.3549086.

[107] R. Angell, B. Johnson, Y. Brun, and A. Meliou, "Themis: Automatically testing software for discrimination," in *Proceedings of the 2018 26th ACM Joint meeting on european software engineering conference and symposium on the foundations of software engineering*, 2018, pp. 871–875.

[108] G. Salvendy, Ed., *Handbook of Human Factors and Ergonomics*, Fourth. Hoboken, NJ: Wiley, 2012, ISBN: 978-0-470-52838-9.

[109] J. C. Bastien, "Usability testing: A review of some methodological and technical aspects of the method," *International Journal of Medical Informatics*, vol. 79, no. 4, e18–e23, Apr. 2010, ISSN: 13865056. DOI: 10.1016/j.ijmedinf.2008.12.004. (visited on 02/02/2023).

[110] T. Gillespie, "Content moderation, AI, and the question of scale," *Big Data & Society*, vol. 7, no. 2, p. 2 053 951 720 943 234, Jul. 2020, ISSN: 2053-9517. DOI: 10.1177/2053951720943234. (visited on 02/15/2023).

[111] S. L. Myers and N. Grant, "Combating Disinformation Wanes at Social Media Giants," *The New York Times*, Feb. 2023, ISSN: 0362-4331. (visited on 02/21/2023).

[112] Social Web Working Group, *ActivityPub*, Jan. 2018. [Online]. Available: https://w3c.github.io/activitypub/%5C#security-considerations (visited on 01/02/2023).

[113] J. H. Saltzer, D. P. Reed, and D. D. Clark, "End-to-end arguments in system design," *ACM Transactions on Computer Systems (TOCS)*, vol. 2, no. 4, pp. 277–288, 1984.

[114] E. Kalliamvakou, G. Gousios, K. Blincoe, L. Singer, D. M. German, and D. Damian, "The promises and perils of mining GitHub," in *Proceedings of the 11th Working Conference on Mining Software Repositories*, ser. MSR 2014, New York, NY, USA: Association for Computing Machinery, May 2014, pp. 92–101, ISBN: 978-1-4503-2863-0. DOI: 10.1145/2597073.2597074. (visited on 03/03/2022).

[115] J. Aranda and G. Venolia, “The secret life of bugs: Going past the errors and omissions in software repositories,” in *2009 IEEE 31st International Conference on Software Engineering*, May 2009, pp. 298–308. DOI: 10.1109/ICSE.2009.5070530.

A. DATA AVAILABILITY

Data is available for access via an artifact (https://zenodo.org/record/7601293) and includes the GitHub issue mining tool and collected data for the entire study.

The issue mining tool was run from the terminal and contains features to interact with the GitHub API and process data from local files. It can download issues and their comments from a project, filter them based on keyword lists, randomize their ordering, and query them after annotation to assist with analysis.

Research data including the baseline keywords (§4.2.1), the keyword tailoring process (§4.2.2), issue sampling (§4.2.3), discussion modeling (§4.3.2), taxonomy development with the codebook (§4.3.3), and inter-rater agreement (§4.3.4) is available. See chapter B for more detail on the codebook.

The documentation in the artifact contains more detail on how to reproduce results for this study.

B. CODEBOOK

The codebooks for the issue type, issue result, and discussion model are provided here. See the artifact (https://zenodo.org/record/7601293) for the full codebook.

Table B.1. Codebook for the issue type.

Label	Description	Example
Bug	A mistake in implementation that deviates from the original design intent.	"Can't suspend users with + sign in their email address" (Mastodon #10576)
Feature request	A proposal for a new addition or modification to the system.	"Instance Greylisting" (Mastodon #4296)

Table B.2. Codebook for the issue result.

Label	Description
Open	Issue is still in the "open" state and is unresolved.
No action	Issue is closed but no change was made to the codebase.
Merged	Issue is closed with some change to the codebase.

Table B.3. Codebook for the discussion model.

Label	**Description**	**Criteria**	**Taxonomy**
Risk	Claim of potential loss that users face when harmed by other users.	Specifically mentions a type of online abuse (e.g. harassment), a scenario that could lead to online abuse, or a weakness that leaves users open to abuse. Reiterated items are not re-coded.	Table 5.2
Option	Proposal to progress the issue towards closure.	Implementation details or UI design are not re-coded.	Table 5.3
Chosen	An option that is selected by engineers.	If maintainers choose the associated option and close the issue, this code should be filled in.	—
Treatment selection rationale	Reason to select an option.	Specifies why a particular option should be selected and acted upon. Only coded for options that are marked as *chosen*.	Table 5.4

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346 USA