Engineering Patterns for Trust and Safety on Social Media Platforms: A Case Study of Mastodon and Diaspora

Geoffrey Cramer[a], William P. Maxam III[b] and James C. Davis[a,*]

[a]Elmore Family School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA
[b]Electrical Engineering and Cyber Systems, US Coast Guard Academy, New London, CT, USA

ARTICLE INFO

Keywords:
Empirical software engineering
Social media platforms
Trust & Safety engineering
Engineering decision-making
Risk

ABSTRACT

Context: Trust & Safety (T&S) Engineering is an emerging area of software engineering that mitigates the risks of harmful interactions in online platforms. Numerous studies have explored T&S risks on social media platforms, taxonomizing threats and investigating individual issues. However, there is limited empirical knowledge about engineering efforts to promote T&S.

Methods: This study examines T&S risks and the engineering patterns to resolve them. We conducted a case study of the two largest decentralised SMPs: Mastodon and Diaspora. These SMPs are open-source, so we analyzed T&S discussions within 60 GitHub issues. We analyzed T&S discussions that took place in their online repository and extracted T&S risks, T&S engineering patterns, and resolution rationales considered by the engineers. We integrate our findings by mapping T&S engineering patterns onto a general model of SMPs, to give SMP engineers a systematic understanding of their T&S risk treatment options.

Results: T&S issues are a challenge throughout the feature set and lifespan of an SMP. A taxonomy of 12 solution patterns are developed, paving the way for academia and industry to standardize Trust & Safety solutions. We conclude with future directions to study and improve T&S Engineering, spanning software design, decision-making, and validation. We 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]. SMPs enable users to share information, express opinions, and be entertained [2], among other benefits [3, 4]. 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 to mitigate these harms [8], 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]. GitHub, a major supporter of open-source and commercial software development, defines T&S *Engineering* as "software designed with user safety in mind" [10]. With a better understanding of how SMPs can be designed to promote trust and safety, software engineers can improve human interactions worldwide. Researchers have previously investigated SMP problems [11, 12, 13] and potential solutions [14, 15, 16, 17]. While prior literature has studied specific interventions, this work aims to characterize the T&S Engineering process itself. Prior work has not elaborated on how engineers identify T&S risks in SMP features, what solutions they explored, and what properties are prioritized in those solutions.

To help address these knowledge gaps, we conducted a *multi-site case study* [18]. We examined T&S work in two SMP contexts, Mastodon and Diaspora. This research method allows us to evaluate T&S Engineering as a phenomenon that has yet to be fully understood as an engineering discipline, although like any case study it raises questions about the generalizability of our data. Our data source was the dialogues associated with T&S-related issues on the studied SMPs. We sampled and analyzed 60 T&S-related issues from two decentralised open-source SMPs, Mastodon (7,833,218 users) and Diaspora (740,409 users) [19]. We sampled T&S issues using keywords, and mapped these examples of the T&S engineering design process onto a discussion model. Finally, we analyzed elements of this discussion model:

*Corresponding author
cramerg@purdue.edu (G. Cramer); william.p.maxam@uscga.edu (W.P.M. III); davisjam@purdue.edu (J.C. Davis)
ORCID(s): 0000-0003-2495-686X (J.C. Davis)

risks, treatments, and rationales. We used a mix of open- and closed- coding to develop taxonomies for T&S risks, engineering patterns, and pattern selection rationales in SMPs. We used inter-rater agreement to validate our results.

Here is a brief summary of our findings. We found that T&S issues continue to occur throughout an SMP's lifespan. Most T&S issues highlight design shortcomings, not implementation errors. T&S issues are difficult to resolve or remain open, with an average resolution time 147 days longer than other issues. We characterized 12 solution patterns for T&S issues, and identified gaps in the current process. When SMP engineers make a design change to improve T&S, their selected treatments are mostly reactive — their preferred approach is to place the burden on moderators ("Add moderation") and users ("Require consent"). When we compare the two case study contexts, we see that many characteristics of T&S issues are similar between Mastodon and Diaspora. However, the Mastodon community is more concerned about T&S risks related to toxic content, while Diaspora is focused on privacy issues. Speaking more broadly, our findings indicate how T&S Engineering is understood by stakeholders, what problems it addresses, and how its efforts are evaluated. In future work, these results can be leveraged to develop more reproducible and generalizable methods that can extend to other SMP contexts.

Our contributions are:

- We describe the first study of T&S Engineering from a software engineering perspective, framed in terms of safety-by-design and the engineering patterns used to effect it (§3).
- We extend taxonomies of T&S risks and threat actors (Table 10), T&S engineering patterns (Table 11), and T&S engineering decision rationales (Table 12), adapting prior work to the T&S engineering context and providing novel measurements of these taxonomies in the SMPs we studied.
- We relate these T&S engineering patterns to the contexts in which they operate (Figure 6) and to the broader principles of T&S by design proposed in prior work Figure 7.
- We share 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 (§9).

2. Background

This section covers the various types of SMPs (§2.1), and discuss the history and role of T&S engineering (§2.2).

2.1. Social Media Platforms

2.1.1. Generalized Framework of SMPs

Many types of online platforms are in use today (SMPs, online markets, messaging platforms, etc.). SMPs are the most popular type of online platform on the Internet; over half of the 20 most visited websites are SMPs [20] and almost 60% of the global population use them [1]. Hopkins defines the many forms of SMPs [21] comprehensively: "Internet-based... and persistent channel[s] of mass personal communication facilitating perceptions of interactions among users, deriving value primarily from user-generated content" [22]. Smith divides SMPs into seven building blocks: identity, presence, relationships, reputation, groups, conversations, and sharing [23], each requiring design [24]. These concepts take many forms in SMPs. For example, "user-generated content" can be hypertext (*e.g.,* Facebook), video (*e.g.,* YouTube), photographs (*e.g.,* Instagram), or records of interactions (*e.g.,* upvotes). SMPs often support more than one type of content.

Figure 1 depicts a context diagram that illustrates the factors at play for T&S engineers. It was developed from the following sources:

- The *Features* element was derived from the building blocks presented by Smith [23].
- The *Moderation* and *Policy* elements were added based on the summary by Singhal *et al.* [8].
- The *Filters* element was developed based on the summary by Singhal *et al.* [8] as well as the initial memoization step of the pilot study (Appendix A).
- The interaction timeline between Alice and Bob was developed based on knowledge from all of the above sources.

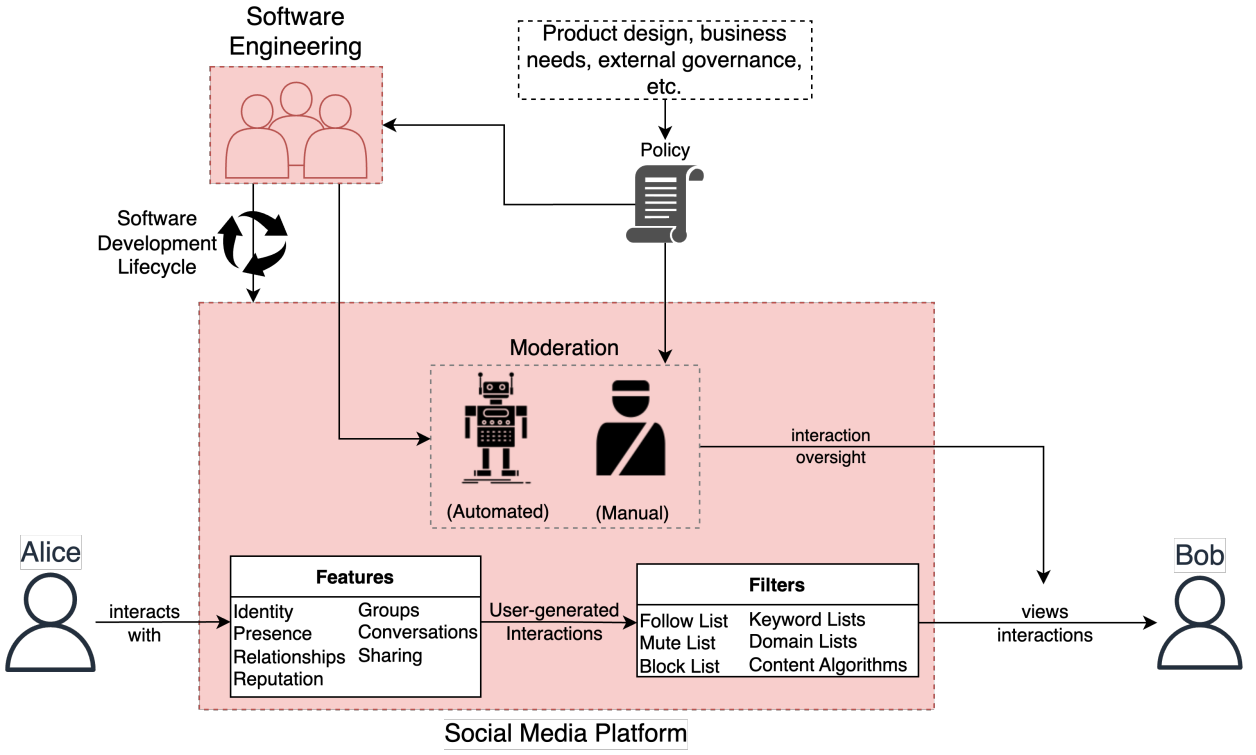


Figure 1: SMP context diagram showing an interaction from Alice to Bob. Eve interacts with features. Her interactions pass through filters and moderation oversight before reaching Bob. Our study's focus is highlighted in pink.

2.1.2. SMP Architectures

With regards to architecture, SMPs can be grouped into two categories, namely centralized and decentralized SMPs. A centralized SMP typically centralizes user visibility, user data, and platform governance into the hands of a single platform operator. All user accounts can interact with one another, possibly influenced by controls on the visibility of an account. The platform operator stores all data on the platform. The platform operator can set and enforce its rules, *e.g.,* moderation.

In contrast, a decentralized SMP decentralizes the visibility of users, the storage of their data, and the governance of the platform. A platform operator ("administrator") deploys an *SMP instance* on a server for public or private use. Each instance functions autonomously, with content policies and moderation defined by the administrator. Users on an instance can see one another, but may not be able to interact with users on other instances. User data may likewise be stored on a per-instance basis, or even represented as hyperlinks to servers operated by the users themselves. Content can be shared across instances through activity stream protocols [27], creating what is called the federated universe or "Fediverse" [28].

While decentralized SMPs have fewer users than their centralized counterparts, they still have millions of users [19], and face many of the same T&S challenges. Part of their appeal is that the decentralised design gives users more control over the content they share and receive, as each instance offers a unique community with distinct rules and options. As an example, those looking to avoid what they perceive as hate speech can subscribe to a private instance that aligns with their preferences [29]. For the purpose of our study, decentralized SMPs are particularly attractive because they tend to be open-source and thus publish much of their design process [30, 31]. Decentralized SMPs have recently seen a large influx of users [19], correlated with the purchase of Twitter (now known as X) by Elon Musk and concomitant concerns about decreased platform Trust and Safety.

Table 1
Description of the elements of an SMP, as depicted in Figure 1.

Element(s)	Description	Example
Software Engineer	Software engineers design and develop an SMP, and are accountable for the quality of the software development process and the end product [25].	In the SMPs considered in this study, the software engineering teams are comprised of a small number of core developers without substantial management oversight, business arm, etc.
Policy	Policies establish norms (conventions) or requirements concerning an SMP's purpose and use [26].	In the case of mastodon.social, in the event that sexually explicit or violent media is being posted, it is required to label such content as sensitive. In the case of Facebook users are not allowed to post violent threats against law enforcement officers.
Features	These are the major features that exist on the platform that fit into the taxonomy from Smith *et al.* [23].	On TikTok, the feature that displays a check mark on prominent user profiles belongs to the Reputation category.
Filters	Filters are components of the platform that allow users to control who sees their content and what content they are exposed to.	An example of a filter on Twitter is the setting that prevents users from tagging you in photos.
Moderation	Moderation promotes organized involvement in an online community by instituting guidelines required to foster cooperation and avoid abuse. [8]	Automated or Manual
Instance	In a decentralized SMP, an administrator can set up a fully independent SMP instance on a server for public or private use.	Mastodon.social

Table 2
Categorization of SMPs along the dimensions of architecture (centralized vs. decentralized) and commercialization. Our study focuses on two decentralized non-commercial SMPs.

Commercialization	Decentralized SMPs	Centralized SMPs
Commercial SMPs	Bluesky, Twetch	Facebook, Twitter, TikTok
Non-commercial SMPs	Mastodon, Diaspora	Locket, Lemon8

2.2. A Software Engineering Perspective on Trust & Safety

2.2.1. History

According to [9], discussions of Trust & Safety originated in the financial sector in the 1990s to address issues such as fraudulent activity. Online platform operators wanted 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 [32]) and in terms of their interactions with other users (*e.g.,* not being spammed or exposed to harmful content) [33]. Over time, it became clear that any online platform where users interact is subject to T&S risk. Efforts to promote T&S were initially distributed across teams, making it difficult to consolidate best practices and apply research findings [34, 35]. 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) [35]. Concurrently, academics at Stanford founded the *Trust & Safety Journal* in 2021 [9]. T&S has therefore broadened to encompass

Table 3
Access and *conduct* boundaries are displayed with their associated domain and risks. Figure 2 conveys this separation in diagram form.

Domain	Description	Threat Actor	Risk Examples
Security & Privacy (Access)	Deals with unpermitted access to data.	Person with unpermitted access to user data.	Data Leak Account compromise Social engineering
Trust & Safety (Conduct)	Deals with misconduct of data that users have legitimate access to.	Person that commits misconduct with user data.	Bullying Trolling Stalking Dogpiling

fraudulent activity, harassment, toxicity, stalking, account takeovers, and any other abusive activity users can take on the platform.

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). However, many organizations 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 [36]. GitLab, Cloudflare, and Pinterest advertise T&S Engineering teams [37, 38, 39, 40]. The TSPA job board lists many T&S opportunities calling for software engineering experience [41].

Today, T&S is spread across the areas of policy, engineering, and moderation [42]. Policymakers specify what constitutes acceptable behavior in the online platform and account for both internal and external governance. Meanwhile, engineering teams design the platform to provide the tools and detection necessary to mitigate T&S risk. Finally, moderation teams enforce the established policies by leveraging the tools that engineering teams design including automated and manual approaches [8].

Promoting Trust & Safety in online platforms is an ongoing problem. A global survey found that 48% of people are affected by online hate and harassment [13]. In 2022, as part of a United Nations action, several nations began work to address online abuse [43]. In 2023, the U.S. surgeon general issued a warning about social media usage and mental health among young people [44]. Our empirical characterization of the open-source T&S engineering process gives the groundwork to improve it.

2.2.2. Definitions: T&S vs. Security & Privacy

To properly situate T&S, we compare it to the related goals of Security and Privacy (S&P). Both are non-functional requirements [45], and both T&S and S&P are desirable properties that engineers strive to achieve. However, S&P is focused on *access* to data while T&S focuses on *conduct* with the data [46]. When someone creates data, there are immediate S&P risks that are present if unpermitted access is granted. When that data is shared openly with others in an online platform, the user that created it should feel *safe* from harm while those that encounter it should be able to *trust* that it is legitimate.

Table 3 lists T&S and S&P domains along with their associated risks. Figure 2 illustrates how the access and conduct boundaries can be violated. We note that the definition of T&S used in this study is somewhat imprecise. We take an empirical perspective and observe engineers' behavior with respect to T&S issues on SMPs, inducing our results from data.

3. Related Work

In this section, we cover prior work on T&S Engineering for SMPs. We discuss the T&S risks in SMPs (§3.1), some solutions to these risks (§3.2), and the existing mapping between safety-by-design and T&S engineering patterns (§3.3). At the end of each subsection, the strength of presented work is analyzed to identify its gaps.

3.1. T&S Risks

Many sources have investigated the T&S risks and threats on social media [47, 48, 11, 49, 50]. Hasib provided a taxonomy of risk categories such as traditional information security (*e.g.,* spam, XSS), identity (*e.g.,* phishing, fake

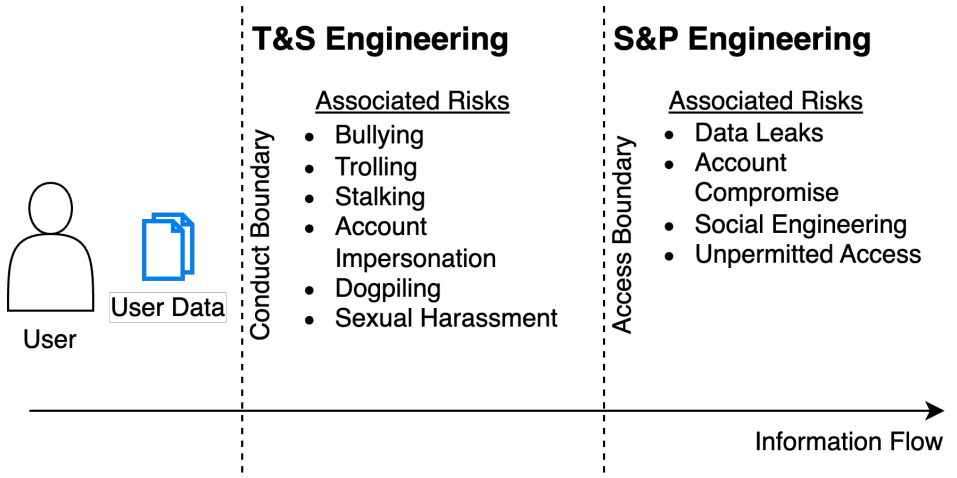


Figure 2: Illustration of *conduct* and *access* boundaries as user data flows through a social media platform. Within the *conduct* boundary are the users that should be able to interact with the data appropriately. Misconduct here yields T&S risks such as account impersonation and trolling. Within the *access* boundary are users that should be allowed to access the data. Here lie S&P risks such as data leaks and account compromises. Table 3 conveys this separation in tabular form.

profiles), privacy (*e.g.,* digital dossiers, facial recognition), and social threats (*e.g.,* stalking) [47]. [48] used a threat modeling approach to SMPs to identify additional threats such as private information disclosure and corporate secrets theft. Other researchers expanded these taxonomies, adding categories such as child-specific threats [11], privacy threats such as deanonymization and location leakage [50, 49], and political threats such as disinformation [51]. [13] provided a recent and exhaustive taxonomy, enumerating myriad forms of online hate and harassment. Due to its recency and sound methods, we view [13] as the state-of-the-art taxonomy of T&S risks. We build on it, identifying two additional categories and extending a third.

Beyond taxonomies, researchers have investigated individual threats. For example: Trabelsi & Bouafif described abuses of content reporting systems [52]; Ashktorab and Vitak investigate cyberbullying mitigation and prevention techniques [5]; [53] analyze social insider attacks; Such *et al.* investigated privacy conflicts in co-owned photos [54]; and Cheng *et al.* studied the efforts of "trolls" to disrupt constructive discussion [55].

Analysis: The basis of these works is strongly grounded. Researchers and practitioners have collaborated to produce a set of risks that are supported by both expert opinion and empirical data.

3.2. T&S Solutions

Two kinds of approaches are used to treat T&S risks on SMPs: design and moderation. Figure 1 illustrates these protection mechanisms. We discuss both here.

Design treatments are *proactive*, preventing T&S issues before they manifest. Some SMP design approaches to promote T&S have been investigated in the literature. A set of solutions from Fire *et al.* [11] include authentication mechanisms, security & privacy settings, internal protection mechanisms, and user reporting features. Prior works have investigated specific design mechanisms to protect users, such as improving authentication and user settings [11], using social relationship-aware content access control [56, 57], experimenting with designs for specific interfaces to prevent abuse [58, 59], and improving moderation interfaces [60, 61, 59]. A recent study proposes changes to SMP architecture to influence end-user behavior, reminding users of platform guidelines before posting certain content [16].

Moderation treatments are *reactive*, limiting the impact of problematic user behaviors after they have occurred (Figure 1). Moderation strategies were summarized by [8] to be human-based, algorithmic-based, or human-in-the-loop. Automated detection techniques have been in use by most major SMPs and can include exact content matching (*e.g.,* using hashes), approximate content matching, and natural language processing [62]. More advanced automation applies deep learning, *e.g.,* convolutional neuron networks, Long-Short-Term-Memory models, and Transformer models [63]. These techniques are inadequate, leaving many users exposed [63].

These prior works have proposed novel solutions to T&S issues. In contrast, our contribution examines what solutions SMP engineers actually apply to address T&S issues and the process they undertake as they do so.

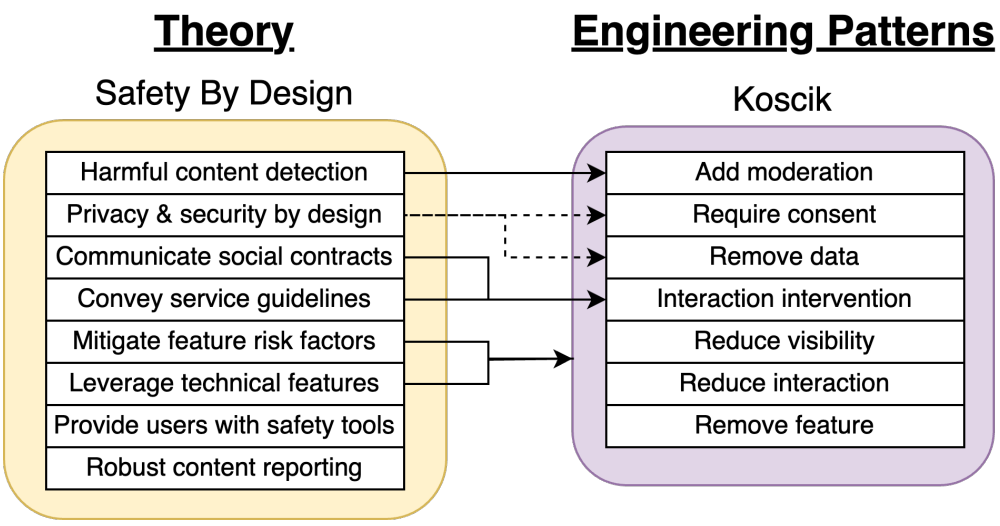


Figure 3: Existing literature that informs *T&S By Design*. The Safety By Design framework [64] on the left provides a theory-based set of guidelines to design safe platforms. [66] (right) lists seven patterns to address abuse vectors. Arrows indicate a relation and dashed arrows indicate a partial relation. The yellow element comes from the Safety By Design framework [64]. The purple element comes from Koscik [66]. Note that our presentation of the Safety By Design list in this figure (left side) is a subset of the original work — we only listed items that involve technical software engineering work. For example, we did not include "develop community guidelines, terms of service and moderation procedures" because this is out of scope for the software engineering focus of our study.

Analysis: The basis of these works is modest and supported by real practice. The solutions listed above have been observed in the field but none of them are exhaustive. More solutions can still be investigated to strengthen the field, especially given the prevalence of T&S risk today in spite of these solutions.

3.3. State-of-the-Art for T&S Engineering

In this section, we discuss the current state-of-the-art for T&S Engineering. To date, there is a gap in operationalizing how software engineers can treat T&S risk. The closest frameworks that attempt systematization are:

1. The *Safety By Design* framework [64]. This framework 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* [65], providing safety tools, leveraging technical features to mitigate risk, evaluating all features to mitigate risk factors, and publishing annual safety assessments.

2. The design pattern taxonomy provided by Koscik [66]. This taxonomy 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 3 shows both prior works and how they inform *T&S By Design*. We drew relations between elements if a suggestion from the *Safety By Design* framework can be implemented with a pattern from Koscik's abuse vector treatment taxonomy. For example, one of the full relations (solid line) is "harmful content detection" and the *Add Moderation* pattern — per Koscik, the only way to detect harmful content with additional moderation. A partial relation (dotted line) is drawn from "privacy & security by design" and Koscik's *Require Consent* and *Remove Data* patterns — both patterns are realizations of the Privacy & Security By Design framework but do not encompass it.

Analysis: The basis of these works is limited with most works coming from industry professional experience and opinion, rather than empirical data. *Safety By Design* is supported by Australia's eSafety Commissioner [64] but does not have available data on how the framework was developed. The design pattern taxonomy [66] appears to be based on Koscik's experience in the field.

Table 4
RQ to method mappings.

RQ	Data Source(s)	Kinds of Analysis
1	Issue feature and type	Closed coding of issue topics
2	Risk statements from discussion model	Qualitative analysis of discussions with subsequent thematic analysis of risk themes
3	Treatment options and rationale statements from discussion model	Qualitative analysis of discussions with subsequent thematic analysis of treatment and rationale themes

3.4. Summary, Unknowns, and Research Questions

To summarize this literature review: 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. Theoretical frameworks exist to describe the concepts of T&S By Design and proposed engineering patterns to achieve this goal. However, we know little of the actual practice of T&S Engineering and of risk-based T&S decision making. Our analysis of prior work shows that they have varying degrees of strength:

1. *T&S risks* have an empirical and professional basis. However, the risk landscape is always changing and could benefit from further data.
2. *T&S solutions* have a moderate grounding in both professionals' experience and real actions taken by SMPs. However, the solution space is infinite and could benefit from empirical data that is readily available to everyone.
3. *T&S Engineering* work is limited with most contributions coming from practitioner experience and little empirical data that is readily available.

The primary gap addressed by this work is this third category: we aim to provide an empirical basis for the work of T&S Engineers when designing SMPs to be safer. Our research provides initial steps toward achieving this goal. We specifically address three research questions (RQs):

- **RQ1**: *What SMP features are affected by T&S?*
- **RQ2**: *What risks are identified in T&S issues?*
- **RQ3**: *What treatment options are proposed in T&S issues? How are they selected?*

Definitions: To scope the broad definition of T&S to our study of SMPs, we define:

- **User T&S** as the study of how users harm other users. 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 SMPs, perhaps because these SMPs lack the profit motivation that leads some commercial platforms to violate T&S in this way. We omitted them during our sampling process.
- **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.

We use **T&S in SMPs** as shorthand for these related concepts.

Contribution: In answering these questions, we obtain an empirically-grounded set of 12 solution patterns (Table 11). In our Discussion section, we examine how our results can help improve the T&S Engineering discipline as a whole (§6.1).

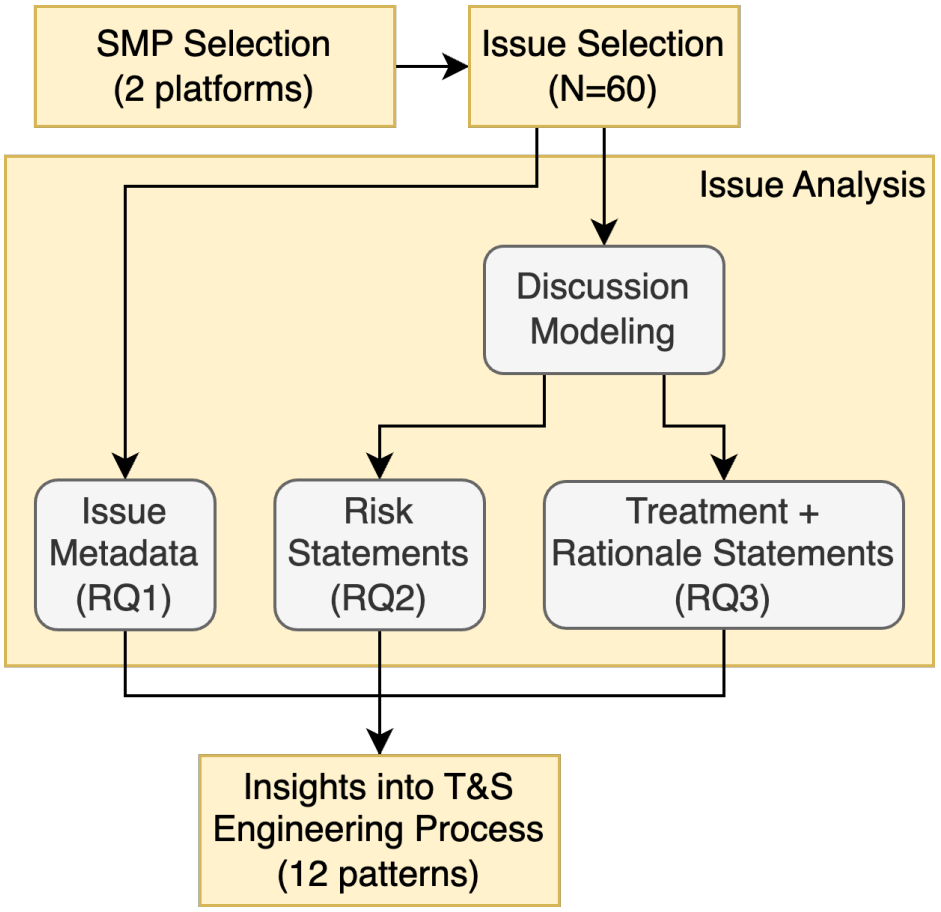


Figure 4: Relationship of research methods and data to RQs. Relationship of research methods and data to RQs. RQ1 is answered with metadata. RQ2 is answered with risk statement data. RQ3 is answered with treatment and rationale statement data.

4. Methodology

To accomplish our goal of studying T&S Engineering, we chose to examine open-source SMPs for a few reasons: all design artifacts are openly available, they have millions of users, and they operate with fewer competing goals unlike their commercial counterparts. We developed a method to select OSS SMP projects (§4.2), filter their T&S discussions (§4.3), and analyze them in a structured manner (§4.4). Using this data, we can answer our research questions. A mapping from research questions to data source and analysis approach is given in Table 4. Figure 4 provides an overview of the methods for this study.

4.1. Summary of Research Team

Author 1 has 6 years of software engineering experience in the physical security industry and personal experience as an SMP user. *Author 2* has 3 years of cybersecurity experience, providing a distinct perspective on how T&S can be improved given that T&S and S&P are closely related (Table 3). *Author 3* has 5 years of software engineering experience and personal experience as an SMP user, as well as 9 years of academic experience that help ensure the work is accurate to the software engineering discipline, is well-designed, and is well-positioned in the literature. These combined experiences allow the study to tackle T&S from a technical perspective (T&S Engineering), shedding light on how this discipline can mature.

Table 5
OSS SMP projects with >100K users. We give the number of users, GitHub issues, and stars as of Jan. 26, 2023. We studied Mastodon and Diaspora, the top two by these metrics.

Project	Category	Users [19]	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

Table 6
SMP filtering results, summarizing resulting keywords, precision and recall in final batch of keyword expansion, number of T&S issues after the selection process, and proportion examined to reach 30 issues per project. We used keywords, aiming for high recall, and addressed the low precision with manual inspection.

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

4.2. SMP Selection

To select the specific SMPs for our study, we consulted an aggregated dataset of such platforms [19]. Table 5 indicates the SMPs tracked by this source (all are decentralized and open-source SMPs). Mastodon and Diaspora are the most popular such platforms, and both have substantial engineering data. We therefore studied Mastodon and Diaspora.

4.3. Issue Selection

Both projects use GitHub and track issues via "GitHub Issues" [30, 31]. In software engineering, an *issue* (*i.e.,* a "ticket", "bug report", etc.) is used to describe a flaw in a system. The reporter (usually a user or an engineer) and the engineering team dialogue about the flaw through a series of posts, similar to a discussion forum. The issue may go unresolved, or be addressed through changes in the system.

We used a keyword approach to find issues associated with these SMPs that contained T&S risk statements. Issue selection followed three phases: selecting baseline keywords, tailoring keywords to the studied projects, and sampling issues. A summary is given in Table 6, and details are shared next. *Author 1* carried out this process with oversight from *Author 3*.

4.3.1. Baseline keywords

To develop a base set of keywords related to T&S, we decided to consult the *Trust & Safety Journal* [67]. At the time of collection, the first two issues of the journal were available. Keywords from both issues were aggregated, then removed/modified based on the following:

- They were removed if they were not directly related to our definition of *T&S in SMPs* (*e.g.,* "robust hashing").
- They were collapsed into a single keyword if they were repeated in different variations (e.g. "content moderation" and "Moderation labor" were combined into "moderation").

In total, *Author 1* removed 43 entries from the list with supervision from *Author 3*. This step resulted in 12 keywords.[1] We used stemming and regular expressions to capture keyword variations. This step reduced Mastodon from 6,523 issues to 659 and Diaspora from 4,699 issues to 182. We applied another filter to ensure adequate discussion: that issues should have ≥5 comments. This reduced Mastodon from 659 → 317 issues and Diaspora from 182 → 113 issues.

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

4.3.2. Keyword Tailoring

Next, we tailored the keyword list to each selected repository. Our goal was to find as many T&S discussions as possible. *Author 1* carried out this process with guidance from *Author 3*. *Author 1* iteratively sampled 100 issues at a time on each of the two platforms. Keywords were added in each round based on the *T&S in SMPs* definition (§3.4). *Author 1* continued until recall >90%. This step expanded Mastodon from 317 issues to 431 and Diaspora from 113 issues to 316.

4.3.3. Issue Sampling

Finally, the issues that matched our keywords and passed our filters were randomly sorted for processing. *Author 1* carried out this step with guidance from *Author 3*. We applied 2 more filters during this step: (1) The issue was relevant based on the *T&S in SMPs* definition (§3.4), and (2) The issue was not marked as a duplicate (this usually means the problem was discussed elsewhere).

While processing issues, we found that issues with many comments were challenging to model (§4.4.1). *Author 1* and *Author 3* agreed to filter out issues with ≥20 comments. This removed a total of 13 issues.

Author 1 then processed issues until a sample size of N=30 was reached in each repository (60 total). This approach weighted equally the T&S issues from each SMP. The stopping point of 60 issues was chosen per resource constraints, but was sufficient to expand the state-of-the-art taxonomy in each dimension we examined. *Author 2* later assisted in measuring inter-rater agreement (§4.4.3).

4.4. Issue Analysis

We iteratively developed an analysis instrument to model the discussions in the issues we sampled. We defined the unit of analysis to be every sentence of every comment in the issue, including the author's initial statement when opening the issue. Details of the process of developing this analysis instrument are given in Appendix A.

4.4.1. Discussion Modeling

Our instrument used a risk-option-rationale model for these discussions. This model, as elaborated in ISO 31000–"Risk Management" [68], considers that an engineering decision requires enumerating treatment options, assessing their associated risks, and giving rationales for choosing among them. Definitions of each element of the model are described next.

Risk We label *risk* identification statements if they contain a T&S risk claim, defined as: the potential loss an SMP faces from users harming other users.

Treatment Option We label risk *treatment option* statements if they advance the issue towards closure (*e.g.,* an implementation idea or proposing "no action").

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

4.4.2. Taxonomization

After modeling each issue, we undertook a round of taxonomic coding for each issue. We started with taxonomies in existing literature, and extended them as needed. Because the final taxonomies are one of our research contributions, we present them in §5. The method for developing the taxonomies was:

- The *SMP feature list* was developed from overarching issue topics and was openly coded (Table 9).
- The *T&S risk taxonomy and threat actor taxonomy* 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 (§4). The threat actor taxonomy was developed using its established definition (An individual or a group posing a threat)[69]) and the basic user roles of SMPs and extended when common actors were identified during coding.
- A *T&S Engineering pattern taxonomy* was developed from *treatment option* statements that treat a T&S risk (Table 11). It extends work from [66]. Annotators started with this taxonomy and iteratively developed new categories for statements that did not fit.

Table 7
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			

- A *rationale taxonomy* was developed from *rationale* statements (Table 12), leveraging existing work [70]. Specifically, we chose the *software quality* taxonomy from [70] because we wanted to capture the desired system properties that influenced decisions.

4.4.3. Analysis Process and Soundness (Inter-rater Agreement)

We modeled each issue with the risk-option-rationale model, sentence by sentence, and discarded sentences that did not fit the instrument. Where we felt it was necessary, two researchers performed the analysis independently, and agreement was reached on cases of disagreement. The level of pre-resolution agreement was measured using Cohen's Kappa coefficient [71].

1. **Initial coding:** *Author 1* coded each issue, developing the codebook using the processes in §4.4.1 and §4.4.2.

2. **Check on the risk-option-rationale discussion modeling:** Following the model in §4.4.1, coded statements from 10% of the T&S issue discussions were provided to *Author 2*, who independently coded the statement. The Kappa coefficients for each type code are: 0.89 for *risk*, 1.0 for *treatment option*, 1.0 for *rationale*, and 1.0 for *chosen* — "almost perfect agreement" for each [72].

3. **Taxonomization:** (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 both analysts to independently code all statements and then resolve disagreements. (3) For the pattern taxonomy, the analysts had a Kappa score of 0.73 ("substantial agreement"). (4) For the rationale taxonomy, a Kappa score of 0.81 was measured ("almost perfect agreement").

4. Based on the high agreement between the independent analysts on most elements of our analysis, we used the single researcher's results for the remaining 90% of the data for all elements except the risk statements (which both analysts coded and reached agreement).

5. Results

5.1. Summary and Examples of Data

First, to illustrate the type of data we collected, we present example T&S issues from Mastodon and Diaspora, coded according to the discussion model described in §4.4.1. Table 7 and Table 8 show these examples.

Table 8
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	

Next, we performed a metadata analysis of T&S issues to summarize their characteristics. Specifically, we determine when they arise and how they are resolved. Figure 5a 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.

Last, we consider T&S issue status and age to get a sense of how effective these engineering teams' T&S risk treatment process is. Examining Figure 5b, we see that: (1) more than one-third of T&S issues are still open with no resolution, (2) closed T&S issues took almost 5 months longer to resolve than the average, (3) Diaspora has closed issues with *no action* much more frequently than Mastodon.

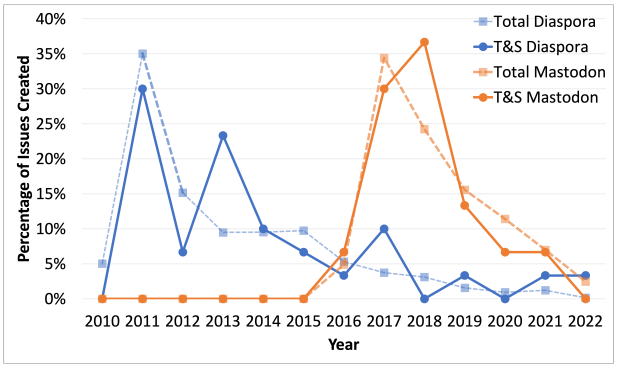
5.2. RQ1: What SMP features are affected by T&S?

Table 9 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*) [23]. Note that we added the *infrastructure* element to this taxonomy, to account for internal features that users do not interact with.

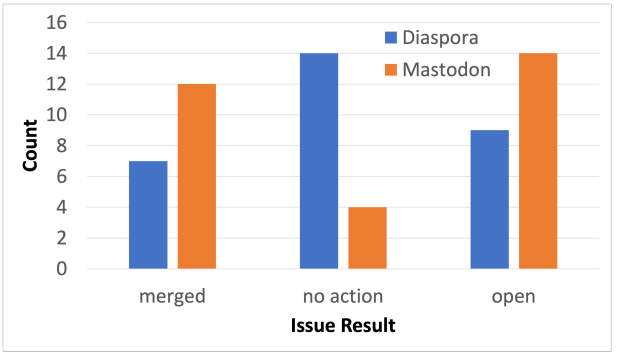
There are some noteworthy differences by platform in each feature's T&S involvement over time. One year after Diaspora's creation, there were a significant number of *content sharing* T&S issues, indicating that this feature posed many T&S risks to the system. In contrast, the early T&S concerns in Mastodon 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.

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

We analyzed the *threat actor* that each risk statement implicated. 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:



(a) Proportion of issues created over time, by SMP (orange–Mastodon; blue–Diaspora) and by type (solid–sampled T&S issues; dashed–all issues). 92% of the studied T&S issues were feature requests, rather than bugs.



(b) 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.

Figure 5: A combined figure of issues over time and their results.

Table 9
SMP feature list. Features were open coded after all T&S issues were gathered (cf. our SMP model in Figure 1). Determinations were based on the primary functionality involved in the issue discussion. Element taxonomy follows Smith [23], with additions in bold.

Feature	Element(s) [23]	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

- *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)

We also carried out the *risk identification* step of the ISO risk management process [68]. Table 10 displays the risk taxonomy and frequencies across 137 risk statements. *Toxic content* is of particular interest in Mastodon, while

Table 10
T&S risks identified in each repository. Taxonomy adapted from Thomas *et al.* [13] with additions in **bold**. Although there are many under "Other", we did not identify any common risk suitable for addition to the taxonomy. Examples of "Other" include issues about blocked and muted instances, and private message deletion/editing.

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
False reporting	Use of content reporting system with malintent.	6	6	12
Impersonation and 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
Other	Risks that do not fit into any other category.	5	11	16

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.

5.4. RQ3: What treatment options are proposed in T&S issues? How are they selected?

To understand the *risk treatment* process, we identify treatment patterns, rationales of treatment selections, and assess the effectiveness of the process itself.

To study treatment patterns, we performed thematic coding on *treatment options* that treat T&S issues. We term these overarching themes as *T&S Engineering patterns*. The initial taxonomy was adopted from Koscik [66] and extended in this study (Table 11).

First, we consider the options that were proposed by discussion members. Table 11 displays each pattern and the frequencies. *Add moderation* is the most frequently proposed pattern, followed by *require consent*. 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.

Table 11
T&S Engineering Patterns. Patterns were coded from *treatment option* statements for T&S issues. Parenthesized digits are the total number of occurrences, and un-parenthesized are the number of unique issues each pattern appears in. P/R indicates **P**roactive vs. **R**eactive patterns (*i.e.,* applied before vs. after an interaction might occurr). C/V/B indicates if the pattern protects the user-generated interaction's **C**reator, **V**iewer, or **B**oth. Additions to Koscik's taxonomy are indicated in **bold**. Two of Koscik's patterns, *Reduce interaction* and *Remove feature* occurred rarely in the studied context.

Pattern	Description	Example	P/R	C/V	Proposed	Chosen
Add moderation	Add or improve moderation tools	"User groups with ACL would be great...could have multiple admins and/or a moderation team" (Mastodon #811)	R	B	20 (35)	7 (8)
Require consent	Ask for approval from involved stakeholders	"I think it should be unchecked, for privacy reasons." (Diaspora #4343)	P	V	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	V	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	B	8 (12)	1 (1)
Improve registration	Bolster user trustworthiness checks	"About spam, what about a captcha during registration?" (Diaspora #4616)	P	V	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	B	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	B	3 (5)	1 (1)
Moderation transparency	Increase clarity of moderation decisions	"I suggest adding [moderation] information to the About page for the instance...people who are vulnerable...can view [it] before deciding on an instance." (Mastodon #8557)	R	B	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	B	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	B	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 [are not] threaded" (Mastodon #1669)	P	B	2 (2)	0 (0)
Remove feature	Take out feature	*Not observed*	P	B	0 (0)	0 (0)

Based on the definitions of each pattern, we refine the previous context diagram (Figure 1) to the point at which each pattern intervenes, and distinguish the actor that each pattern relies on. The new context diagram is shown in Figure 6. The diagram is split by a few dimensions:

- *Risk Scenario*: In the top diagram, a benign user, Alice creates content that could potentially be viewed by a bad actor. In the bottom diagram, a bad actor creates content that a benign user, Bob, could potentially view. Patterns in the middle can be used for both scenarios.
- *Proactive/Reactive*: The dashed line indicates whether a pattern *proactively* intervenes before an interaction occurs, or *reactively* intervenes afterward.
- *Pattern Type:* Additionally, we signify in color the party on which each pattern relies. Seven of the identified patterns are *proactive* in nature, while five are *reactive*. Four patterns rely on humans, while the other eight are automated.

Next, we analyze what patterns are actually chosen by engineers and why. Figure 6 shows that *proactive* patterns are chosen less frequently by engineers. The selected options typically rely on users or moderators rather than automation. Table 12 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 rely on human intervention. 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).

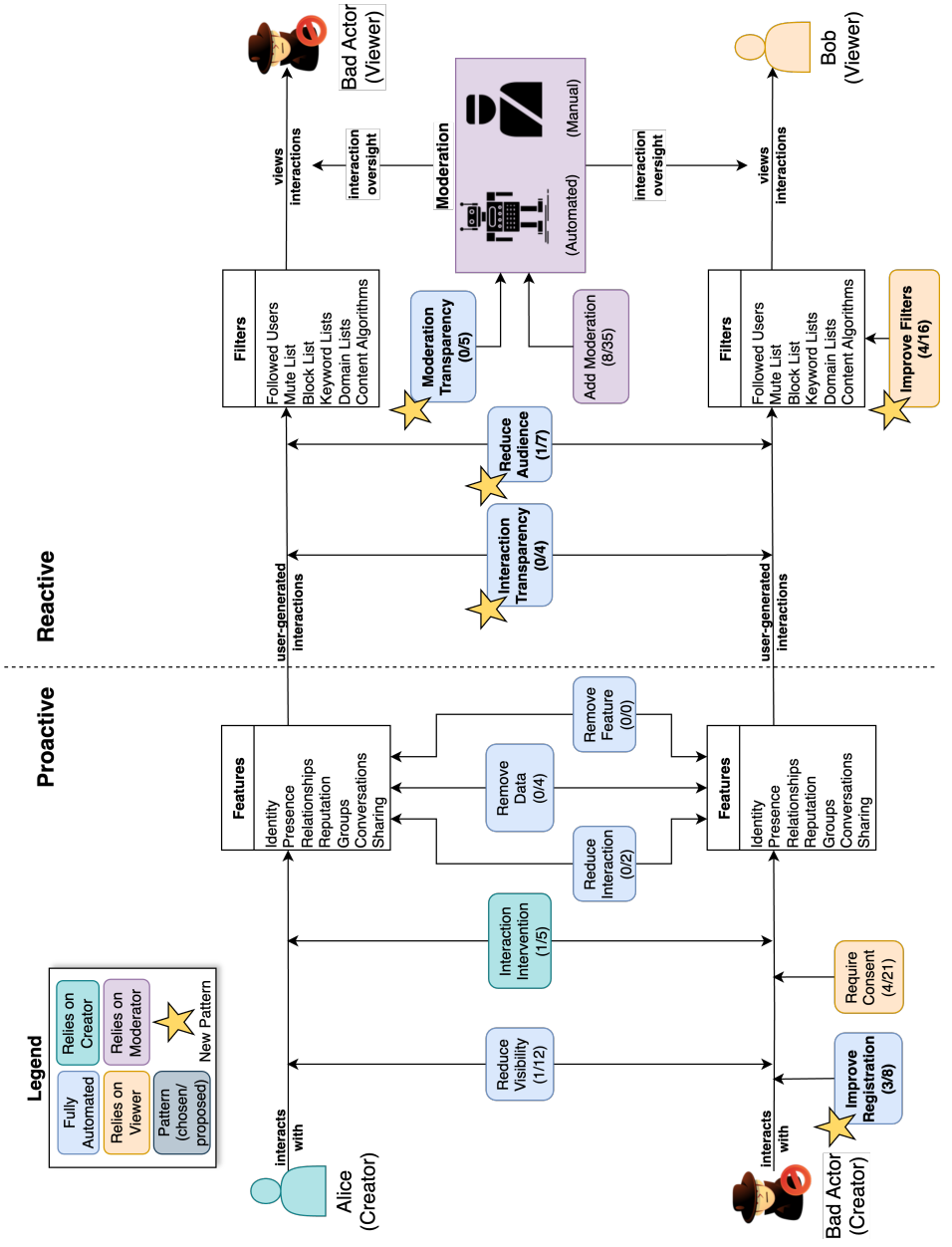


Figure 6: SMP context diagram with T&S Engineering patterns. Two scenarios are shown. The top scenario shows a benign user, Alice, creating content that could be viewed by a bad actor. The bottom scenarios depicts a bad actor creating problematic content to which a benign user, Bob, could be exposed. Some of the patterns are suitable for either context, while others are specific to the bottom scenario. Patterns that intervene to the left of the dashed line are *proactive* (taking place before the content becomes visible) and those to the right are *reactive* (used to mitigate the effect of problematic content or behavior). See the legend for more detail.

Table 12
T&S risk treatment rationales, distinguished by their frequency of mention in issues for which a solution was merged vs. no action was taken. We contextualize the taxonomy from Ko *et al.* to T&S [70]. **Bold**: new categories.

Result	Rationale	Description	Example	Count
MERGED	**Safety**	Protects user from T&S risks	"The primary motivator would be that I just don't want to see Bad Person's bad posts while browsing in-app." (Mastodon #7741)	14
	Moderator efficiency	Allows moderators to complete desired actions	"Of course moderators can decide not to use direct messages, but moderation in the open is mostly not very productive." (Mastodon #8969)	9
	Feasibility	Ease of implementation	"So, populating the checklist shouldn't be hard." (Mastodon #423)	6
	Flexibility	Handles a variety of use cases	"It could be a great compromise between letting users do all what they want and deleting their accounts once for all." (Diaspora #5564)	6
	Clarity	Provides clear experience to users	"This is because in its current iteration: that is what it is, a 'hard mute'." (Mastodon #231)	4
	Security	Prevents unwanted data access	"It is an issue that admins can access unflagged private/direct messages." (Mastodon #6986)	3
	User efficiency	Allows user to easily complete desired action	"Currently one needs to go to that particular post by clicking on the time stamp." (Diaspora #1667)	3
	Annoyance	Removes unnecessary hindrance to user activity	"It means a million extra clicks...to interact with the thread." (Mastodon #1123)	2
NO ACTION	**Unsafety**	Adverse effect to user T&S	"On Twitter, DMs became a terrible spam vector and links in them were banned to try and mitigate this." (Mastodon #90)	9
	Infeasibility	Difficulty of implementation	"I'm not aware of a technical possibility to prevent [unpermitted access] in a distributed network." (Diaspora #3863)	9
	Federation incompatibility	Not possible due to sharing protocols	"This is technically very difficult to do right now in a federated manner, because we don't support editing" (Diaspora #2121)	6
	Insecure	Susceptible to unwanted data access	"This is to avoid accidental leak of a private post to an unwanted recipient and makes the federation protocol a lot easier as a side effect." (Diaspora #6596)	5
	Inconsistency	Conflicts with design or user expectations	"If it's been read already, then it's mutual property" (Diaspora #1828)	3
	Uncertainty	Unclear design or T&S environment	"That's a good point, and we'll probably revisit this in a week or so to see how people are using it." (Diaspora #1369)	1
	Annoyance	Adds unnecessary hindrance to user activity	"Every user has to subscribe to the shared blocklist." (Mastodon #1092)	1
	Unclarity	Complicated or convoluted user experience	"Mastodon aims to be useable by 'non-tech-savvy' people (I guess that implies basic 'online safety' measures as well)" (Mastodon #8340)	1

6. Discussion and Future Work

Here, we first examine the direct implications for the studied (decentralized, non-commercial) SMPs based on our findings (§6.1). Then, we sketch the generalizability of our results by examining one T&S risk in a centralized commercial SMP, TikTok (§6.2). Finally, we describe directions for future work (§6.3).

6.1. Recommendations for SMPs

Based on our findings, we suggest several ways in which SMPs might improve their T&S risk management process.

6.1.1. Communicate Existing Risks.

Sec. 5.4 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. Complementing this concern, treatment patterns that add transparency to the SMP (viz. 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. These findings are not surprising. T&S Risks are present in many forms [11, 50] and have worldwide impact [13].

6.1.2. Document risk sources and treatment options.

In the SMPs we studied, the knowledge of risky features, risk factors, and risk treatments is distributed across project personnel and documents (*e.g.,* distinct discussions scattered across many similar issues). We identified patterns within these features (Table 9), factors (Table 10), treatments (Table 11), and rationales (Table 12). In particular, the *add moderation* treatment pattern is discussed very frequently, yet was selected in only 7/20 issues (Table 11). Given that to date, automated moderation in OSS SMPs is limited, these suggestions would simply place more load on human moderators. It is well-known that content reporting mechanisms themselves can be abused [52], indicating that this solution (in the absence of automated moderation) only shifts risks from end users to human moderators. A clearer understanding of typical solutions would allow these discussions to be more productive, explore the solution space more effectively, and could first opt for automated solutions before exploring to manual ones. These patterns can also 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.

6.1.3. Formalize T&S Reporting and use it as Engineering Feedback

To the best of our knowledge, the studied 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 [73] and Pew Research [6], first-party data will provide clarity into the current T&S risk landscape. Methods for these 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 [74], Snapchat [75], and Discord [76] 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. A similar approach for open-source SMPs could be beneficial. We acknowledge, however, the potential complexity of collecting such data in a distributed SMP. Another challenge is triaging the T&S issues at the scale of large SMPs; the recent proposal by Anandayuvaraj *et al.* to use large language models to filter issues and automatically build postmortems [77] might be helpful.

Our examination of the T&S defect arrival rate (Figure 5a) 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 [33], the Trust & Safety Foundation [73], and the Trust & Safety Journal [9] exist because promoting T&S is a complex, endless pursuit. It requires effort from many stakeholders, including engineers.

6.1.4. Explore proactive solutions.

In our data, we examined 60 issues of which 19 were resolved with a change. As illustrated in Figure 6, most of these solution approaches were reactive rather than proactive. They generally *shared risk* between users and moderators of the system. Can SMPs pursue proactive patterns instead to prevent T&S risks before they are realized?

6.2. Connecting to Commercial SMPs: TikTok's Handling of Young Users

One criticism of our work is the focus on distributed, non-commercial, open-source SMPs. This focus may limit the generalizability of our findings to commercial SMPs. To understand the potential generalizability of our findings, we

examined one T&S issue for the TikTok SMP. We describe the issue timeline and connect it to our findings. We show that (1) all of the actions can be described using our taxonomies, and that (2) our T&S engineering patterns suggest additional approaches beyond those taken by TikTok.

Background: TikTok has enforced a minimum age for users based on the laws of each country it operates in [78]. 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 [79]. 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 [80]. 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 [81]. 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 [82]. 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 [83]. In April 2023, TikTok was fined £12.7 million by the U.K.'s Information Commissioner's Office for misusing data of young users [84].

TikTok's responses: How has TikTok responded to these critical societal threats? We consulted news releases with the "Safety" tag from the TikTok newsroom page [85] 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 [86]. Later in 2019, TikTok announced the "TikTok for Younger Users" feature that limits sharing, comments, and other interactions [87]. 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 [88]. 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 [89, 90]. 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 [91]. In late 2021, TikTok also posted that new educational resources are made available to parents as part of their "Family Pairing" feature [92]. 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.

Analysis: We frame TikTok's responses in terms of the T&S engineering patterns we observed in Mastodon and Diaspora (cf. Table 11). Table 13 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. TikTok then began developing customized features 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. 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. 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 transparency* by publishing routine reports.

Table 13
How TikTok has protected young users. Actions span across the *moderation*, *content sharing*, *user registration*, *private messaging*, *content filtering*, *and user filtering* features. Patterns refer to Table 11.

Feature	Year	Action	Pattern(s)
Moderation	2019	Report content or a profile directly from within the app [93]	add moderation
	2021	Underage account takedowns [89]	add moderation
	2021	Share information regarding removals of suspected underage accounts [94]	moderation transparency
Content Sharing	2019	Allow comments from followers only [95]	require consent
	2019	Control who can duet or react to videos [91]	require consent
	2019	Disable sharing/commenting if under 13 [96]	reduce visibility
	2021	Restriction of TikTok LIVE for young users [97]	reduce visibility
	2021	Make account private [98]	require consent
	2021	Disable Duet and Stitch if under 16 [91]	reduce visibility
	2021	Disable downloads on content from accounts under 16 [91]	reduce visibility
	2021	Pop-up when teenagers first post to choose visibility level [91]	interaction intervention
User Registration	2019	Enforce age-appropriate experiences [91]	reduce visibility, reduce interaction
	2021	Require manual entry of full birthdate [90]	interaction intervention
Private Messaging	2019	Teens can only get messages from followers [91]	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 [88]	reduce visibility
	2021	Default direct messaging setting to 'no one' for ages 16-17 [91]	reduce visibility
Content Filters	2019	Enable comment filters [95]	improve filters
	2021	No notifications after 9pm if age 13-15 [91]	reduce visibility
	2021	No notifications after 10pm if age 16-17 [91]	reduce visibility
	2023	Forced 60-minute time limit if under 18 [99]	reduce visibility
User Filters	2019	Removing unwanted followers [100]	require consent
	2019	Block unwanted users [101]	require consent

By leveraging the model of T&S engineering patterns shown in Figure 6, TikTok's existing strategies can be characterized and additional possible approaches can be considered. For example, 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 (intent on subverting the age policy). 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

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

6.3.1. T&S Engineering Pattern Catalog

Tables 10 and 11 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 SMPs (Table 5), could improve this catalog. The T&S risks on commercial SMPs could also be incorporated, *e.g.,* following the method of [102], although the solution patterns are sometimes opaque. Figure 6 is a starting point for such work.

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

6.3.2. Improved T&S Testing

We note a surprising *absence* in our codebook: T&S testing, *i.e.,* validation that an SMP feature does not contain T&S risks and/or validation of a T&S mitigation. Operationalizing T&S for automated testing is an open challenge. However, due to the contextual nature of T&S risks, fully automated techniques such as those of [103] and [104] could be limited. For example, automated T&S testing could check that basic user boundaries are respected, but this would require models for normal and abnormal user behavior, for user boundaries, for consent, and so on. A possible starting point for this research direction is the usability testing literature [105, 106].

6.3.3. Automated Content Moderation in Federated, Non-Commercial SMPs

Commercial SMPs rely heavily on automated moderation, while SMPs tend to use human moderation. Human moderation has limits — *under moderation* is a frequent T&S risk in SMPs (Table 10). However, developing accurate automated moderation has proven challenging because of the amount of contextual information required to make a judgment. To what extent can automated content moderation be incorporated into SMPs? Is a decentralized 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, non-commercial SMP stakeholders may be unwilling to adopt automated content moderation due to highly-publicized failures in commercial SMPs, and due to the high costs of training and operating the machine learning models that underpin automated content moderation. Understanding these human factors and the effect of (non-)commercialization could advance the conversation.

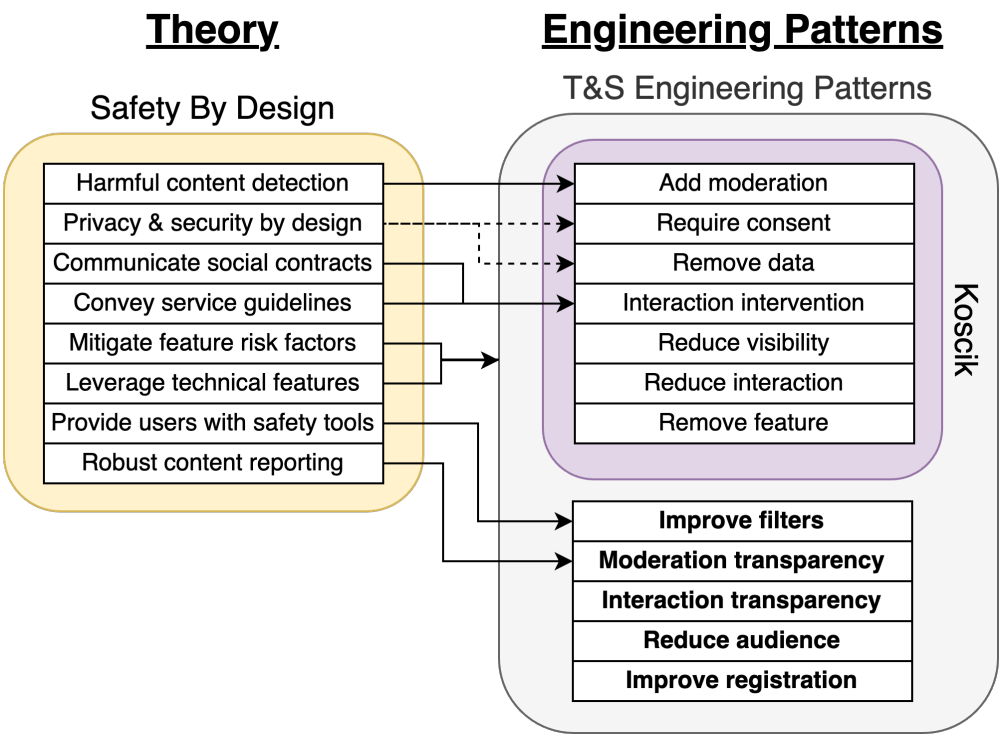


Figure 7: Contributions of this study that inform *T&S By Design*. This study builds upon Koscik [66] while reinforcing the Safety By Design framework [64]. This figure extends Figure 3 and adds the gray element which encapsulates all known T&S Engineering patterns.

6.3.4. T&S Improvements in Federated Protocols

Federation incompatibility was cited in 7 proposal rejections (Table 12). Thus, federation protocols expose SMPs to substantial risk. Adding safety features within the protocol (*e.g.,* anti-spam measures [107]) 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 [108], but perhaps some improvement is possible.

6.3.5. T&S By Design

As discussed in §6.1.4, many of the T&S engineering patterns we observed were reactive, addressing T&S issues by intercepting problematic behavior or content after it has been generated. Prior works have studied how T&S can be incorporated into comment thread design [58] and SMP design [16], but as yet there is no general agenda for T&S by Design. This direction should be informed by fields such as *Privacy by Design* [65, 109] and *Security by Design* [65, 110]. Rubinstein & Good argue that past SMP privacy failures could be avoided through a design approach [111]. Leveraging the current study to inform T&S engineering processes may allow engineers to move from re-actively improving T&S to proactively promoting T&S by design.

As discussed in §3.3, the closest effort we have to an encompassing T&S design process is the abuse vector mitigation strategies from Koscik [66] and the general principles provided by the *Safety By Design* framework [64]. This study builds upon these by providing an empirical basis and adding new design patterns (Table 11). Figure 7 shows how this study contributes to these prior works. By studying safe software design, practitioners can develop measures that are effective [58, 16], scalable, and preventative. The proactive treatment patterns from Table 11 provide a starting point for such work.

7. Threats to Validity

Internal validity Our methodological choices that could affect our findings. *First*, our work relied on qualitative analysis. To reduce bias, we measured inter-rater agreement. To promote comparisons across studies, we used

existing taxonomies, extending them as needed. *Second*, our work mined GitHub. This carries concomitant general concerns [112, 113]. There is also a Diaspora-specific concern. Diaspora uses a separate forum to discuss preliminary feature proposals [31]. Some of these proposals are subsequently filed on GitHub; we only studied such. This data source was omitted because those proposals do not include actions taken by OSS engineers.

External validity The primary threat to this work is its generalizability. We 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). We note two mitigating features of our work. First, although the SMPs we studied are a fraction of the size of SMPs such as Facebook, they nevertheless have over 8 million users — T&S concerns affecting 8 million users are worth studying. Second, although we studied open-source decentralized SMPs, we built our analysis on top of existing taxonomies derived from commercial SMPs. Our data fit these taxonomies, suggesting similarities between the commercial and non-commercial contexts, although in each case we observed new behaviors that required extending the taxonomies.

As a secondary concern, we studied only N=60 issues, 30 from each SMP. A larger sample size could increase the scope of our findings. We note that we analyzed 73% of Diaspora issues (Table 6), indicating that the data was approaching exhaustion for that project. Furthermore, even within this sample, we were able to extend each existing taxonomy that we applied.

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 our data. We 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. We then analyzed those issues using our own understanding of T&S risks by leveraging an ISO risk management standard [68]. However, there is no guarantee that the OSS engineers were using the same terminology. We mitigated this by measuring information retrieval on our keywords.

8. Conclusion

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 SMPś T&S risks.

We conducted the first empirical study of T&S risks on SMPs from a software engineering perspective. We studied 60 T&S-related GitHub Issues for the two most popular open-source SMPs, Mastodon and Diaspora. Our work identified novel SMP risks, engineering patterns, and resolution rationales. Our 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 rather than proactive; and (4) Selected treatments mostly share risk with users or moderators, despite many alternatives. Our work suggests that, in open-source SMPs, there is currently no systematic engineering approach to promoting T&S. We show opportunities for research on software design, decision-making, and validation for T&S in SMPs.

9. Data Availability

We share replication data via an anonymized artifact, including codebook, sampled issues, models, and multi-rater codes. See `https://zenodo.org/record/7601293`.

Acknowledgments

We acknowledge the helpful and critical input from the reviewers and from A. Kazerouni, T. Zhang, and especially A. Marwick. We thank A. Tewari, L.Q. Li, S. Krithivasan, E. Gorostiaga Zubizarreta, and C.M. Sale for assistance in preliminary studies and data collection.

A. Pilot Study to Develop an Analysis Instrument

First, *Author 1* performed general issue memoing to get a sense of issue content. Combining this data with related works [114, 115, 116], it was concluded that discussions could be modeled as arguments [117], but they were also risk-oriented. We tried several ways to model them.

- We started with basic argumentation modeling [117] by using the work from [114]. We found that the model was too general and did not capture themes of interest.
- Next, we attempted to model discussions after the generic Twente Argumentation Schema [118]. [115] put this to work in their software engineering work which aimed to determine the effects of uncertainty in software engineering discussions. We found this process was very slow, error-prone, and did not bring us much closer to answering the RQs.
- Next, we decided to layer a risk-oriented framework on top of a simplified version of the Twente Argumentation Schema. We chose ISO's 31000:2018 risk management framework [68] due to its generalizability and wide acceptance. This approach would use a tested framework but also capture risk elements related to our research topic. This approach was more effective at capturing relevant discussion elements. However, we integrated a second analyst to assess the reliability of our modeling (inter-rater agreement) and observed low inter-rater agreement.
- After finding low inter-rater agreement scores, the analysis instrument was simplified to only use the ISO risk management framework. The three elements of the ISO risk management framework that consistently occurred in discussions were: *risk*, *treatment option*, and *treatment selection rationale*. Analysts could detect these reliably, and they captured the data of interest. This simplified model is conceptually consistent with more general theories of argumentation used in software engineering research [114, 115].

References

[1] D. Chaffey, Global social media statistics research summary 2022 (Aug. 2022).

[2] A. Whiting, D. Williams, Why people use social media: A uses and gratifications approach, Qualitative Market Research: An International Journal 16 (4) (2013) 362–369. doi:10.1108/QMR-06-2013-0041.

[3] S. E. Rolland, G. Parmentier, The Benefit of Social Media: Bulletin Board Focus Groups as a Tool for Co-creation, International Journal of Market Research 55 (6) (2013) 809–827. doi:10.2501/IJMR-2013-068.

[4] R. Deloatch, B. P. Bailey, A. Kirlik, 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, CHI '17, Association for Computing Machinery, New York, NY, USA, 2017, pp. 736–747. doi:10.1145/3025453.3025709.

[5] Z. Ashktorab, 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, CHI '16, Association for Computing Machinery, New York, NY, USA, 2016, pp. 3895–3905. doi:10.1145/2858036.2858548.

[6] E. A. Vogels, The State of Online Harassment, Tech. rep., Pew Research Center (Jan. 2021).

[7] A. Marwick, B. Clancy, K. Furl, Far-Right Online Radicalization: A Review of the Literature, The Bulletin of Technology & Public Life (May 2022). doi:10.21428/bfcb0bff.e9492a11.

[8] M. Singhal, C. Ling, P. Paudel, P. Thota, N. Kumarswamy, G. Stringhini, S. Nilizadeh, SoK: Content Moderation in Social Media, from Guidelines to Enforcement, and Research to Practice (Oct. 2022). arXiv:2206.14855, doi:10.48550/arXiv.2206.14855.

[9] E. Cryst, S. Grossman, J. Hancock, A. Stamos, D. Thiel, Introducing the Journal of Online Trust and Safety, Journal of Online Trust and Safety 1 (1) (Oct. 2021).

[10] L. Galantino, Trust & Safety Engineering @ GitHub (May 2019).

[11] M. Fire, R. Goldschmidt, Y. Elovici, Online Social Networks: Threats and Solutions, IEEE Communications Surveys Tutorials 16 (4) (2014) 2019–2036. doi:10.1109/COMST.2014.2321628.

[12] A. M. Memon, S. G. Sharma, S. S. Mohite, 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 60 (4) (2018) 384–392. doi:10.4103/psychiatry.IndianJPsychiatry_414_17.

[13] K. Thomas, D. Akhawe, M. Bailey, D. Boneh, E. Bursztein, S. Consolvo, N. Dell, Z. Durumeric, P. G. Kelley, D. Kumar, D. McCoy, S. Meiklejohn, T. Ristenpart, G. Stringhini, SoK: Hate, Harassment, and the Changing Landscape of Online Abuse, in: 2021 IEEE Symposium on Security and Privacy (SP), 2021, pp. 247–267. doi:10.1109/SP40001.2021.00028.

[14] J. Qian, A. Bethke, Y. Liu, E. Belding, W. Y. Wang, A Benchmark Dataset for Learning to Intervene in Online Hate Speech, arXiv:1909.04251 [cs] (Sep. 2019). arXiv:1909.04251.

[15] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, D. Testuggine, The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes, arXiv:2005.04790 [cs] (Jun. 2020). arXiv:2005.04790.

[16] J. Kim, C. McDonald, P. Meosky, M. Katsaros, T. Tyler, Promoting Online Civility Through Platform Architecture, Journal of Online Trust and Safety 1 (4) (Sep. 2022). doi:10.54501/jots.v1i4.54.

[17] S. Arumugam, 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 11 (1) (2023) 195–207.

[18] P. Ralph, N. bin Ali, S. Baltes, D. Bianculli, J. Diaz, Y. Dittrich, N. Ernst, M. Felderer, R. Feldt, A. Filieri, B. B. N. de França, C. A. Furia, G. Gay, N. Gold, D. Graziotin, P. He, R. Hoda, N. Juristo, B. Kitchenham, V. Lenarduzzi, J. Martínez, J. Melegati, D. Mendez, T. Menzies, J. Molleri, D. Pfahl, R. Robbes, D. Russo, N. Saarimäki, F. Sarro, D. Taibi, J. Siegmund, D. Spinellis, M. Staron, K. Stol, M.-A. Storey, D. Taibi, D. Tamburri, M. Torchiano, C. Treude, B. Turhan, X. Wang, S. Vegas, Empirical Standards for Software Engineering Research, arXiv:2010.03525 [cs] (Mar. 2021). arXiv:2010.03525.

[19] The Federation, The Federation - a statistics hub, https://the-federation.info/ (n.d.).

[20] Alexa top 1 million websites (2022).
URL https://www.expireddomains.net/alexa-top-websites

[21] C. T. Carr, R. A. Hayes, Social Media: Defining, Developing, and Divining, Atlantic Journal of Communication 23 (1) (2015) 46–65. doi:10.1080/15456870.2015.972282.

[22] J. Hopkins, How to Define Social Media – An Academic Summary (Oct. 2017).

[23] G. Smith, Social Software Building Blocks, https://web.archive.org/web/20171123070545/http://nform.com/ideas/social-software-building-blocks (Apr. 2007).

[24] J. H. Kietzmann, K. Hermkens, I. P. McCarthy, B. S. Silvestre, Social media? Get serious! Understanding the functional building blocks of social media, Business Horizons 54 (3) (2011) 241–251. doi:10.1016/j.bushor.2011.01.005.

[25] C. Irvine, The software engineer: Role, responsibilities and education (1967).

[26] T. G. Gillespie, The politics of 'platforms' (2010).

[27] Wikipedia, Activity stream, Wikipedia (Sep. 2022).

[28] A. Mansoux, R. Roscam Abbing, Seven theses on the fediverse and the becoming of floss (2020).
URL https://www.diva-portal.org/smash/get/diva2:1699767/FULLTEXT01.pdf

[29] D. Zulli, M. Liu, R. Gehl, Rethinking the "social" in "social media": Insights into topology, abstraction, and scale on the Mastodon social network, New Media & Society 22 (7) (2020) 1188–1205. doi:10.1177/1461444820912533.

[30] mastodon, Mastodon/mastodon, Mastodon (n.d.).

[31] diaspora, Diaspora/diaspora, diaspora (n.d.).

[32] N. Confessore, Cambridge Analytica and Facebook: The Scandal and the Fallout So Far, The New York Times (Apr. 2018).

[33] Trust and Safety Professional Association, What We Do, https://www.tspa.org/what-we-do/ (n.d.).

[34] A. Marwick, Trust & safety: The formalization of profession (March 2021).
URL https://tiara.org/wp-content/uploads/2021/07/amarwick_cv072021.pdf

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

[36] D. Leong, Adding Community & Safety checks to new features (Jan. 2017).

[37] Trust and Safety Professional Association, Senior Security Engineer, Trust & Safety, https://web.archive.org/web/20220808140939/https://www.tspa.org/j security-engineer-trust-safety/ (Jul. 2022).

[38] Cloudflare, Trust & Safety Engineering Team, https://web.archive.org/web/20220128033140/https://www.builtinaustin.com/job/engineer/software-engineer-trust-safety-engineering-team/76783 (Dec. 2021).

[39] M. Samuelson, How Pinterest built its Trust & Safety team (Apr. 2022).

[40] S. Xie, Building a Label-Based Enforcement Pipeline for Trust & Safety (May 2021).

[41] Trust and Safety Professional Association, TSPA Job Board, http://web.archive.org/web/20221219211619/https://www.tspa.org/explore/job-board/ (n.d.).

[42] Trust and Safety Professional Association, Content moderation and operations, httpswww.tspa.org/curriculum/ts-fundamentals/content-moderation-and-operations/what-is-content-moderation/ (Jun. n.d.).

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

[44] HHS Department, Surgeon General Issues New Advisory About Effects Social Media Use Has on Youth Mental Health, https://www.hhs.gov/about/news/2023/05/23/surgeon-general-issues-new-advisory-about-effects-social-media-use-has-youth-mental-health.html (May 2023).

[45] I. Sommerville, Software Engineering, Vol. 137035152, Pearson Education, 2015.

[46] The Digital Trust & Safety Partnership, The safe framework: Tailoring a proportionate approach to assessing digital trust & safety, accessed: 2024-10-28 (December 2021).
URL https://dtspartnership.org/wp-content/uploads/2021/12/DTSP_Safe_Framework.pdf

[47] A. A. Hasib, Threats of online social networks, International Journal of Computer Science and Network Security 9 (11) (2009) 288–293.

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

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

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

[51] Y. Wang, M. McKee, A. Torbica, D. Stuckler, Systematic Literature Review on the Spread of Health-related Misinformation on Social Media, Social Science & Medicine 240 (2019) 112552. doi:10.1016/j.socscimed.2019.112552.

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

[53] W. A. Usmani, D. Marques, I. Beschastnikh, K. Beznosov, T. Guerreiro, L. Carriço, Characterizing Social Insider Attacks on Facebook, in: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, CHI '17, Association for Computing Machinery, New York, NY, USA, 2017, pp. 3810–3820. doi:10.1145/3025453.3025901.

[54] J. M. Such, J. Porter, S. Preibusch, 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, CHI '17, Association for Computing Machinery, New York, NY, USA, 2017, pp. 3821–3832. doi:10.1145/3025453.3025668.

[55] J. Cheng, M. Bernstein, C. Danescu-Niculescu-Mizil, 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, CSCW '17, Association for Computing Machinery, New York, NY, USA, 2017, pp. 1217–1230. doi:10.1145/2998181.2998213.

[56] N. Kashmar, M. Adda, H. Ibrahim, M. Atieh, Access Control in Cybersecurity and Social Media, Access Control in Cybersecurity and Social Media (Feb. 2021).

[57] G. Misra, J. M. Such, How Socially Aware Are Social Media Privacy Controls?, Computer 49 (3) (2016) 96–99. doi:10.1109/MC.2016.83.

[58] J. Seering, T. Fang, L. Damasco, M. C. Chen, L. Sun, 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, CHI '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 1–14. doi:10.1145/3290605.3300836.

[59] K. Mahar, A. X. Zhang, D. Karger, Squadbox: A Tool to Combat Email Harassment Using Friendsourced Moderation, in: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI '18, Association for Computing Machinery, New York, NY, USA, 2018, pp. 1–13. doi:10.1145/3173574.3174160.

[60] S. Jhaver, C. Boylston, D. Yang, A. Bruckman, Evaluating the Effectiveness of Deplatforming as a Moderation Strategy on Twitter, Proceedings of the ACM on Human-Computer Interaction 5 (CSCW2) (2021) 381:1–381:30. doi:10.1145/3479525.

[61] S. Jhaver, S. Ghoshal, A. Bruckman, E. Gilbert, Online Harassment and Content Moderation: The Case of Blocklists, ACM Transactions on Computer-Human Interaction 25 (2) (2018) 1–33. doi:10.1145/3185593.

[62] R. Gorwa, R. Binns, C. Katzenbach, Algorithmic content moderation: Technical and political challenges in the automation of platform governance, Big Data & Society 7 (1) (2020) 2053951719897945. doi:10.1177/2053951719897945.

[63] W. Wang, J.-t. Huang, W. Wu, J. Zhang, Y. Huang, S. Li, P. He, M. R. Lyu, MTTM: Metamorphic Testing for Textual Content Moderation Software, in: 2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE), 2023, pp. 2387–2399. doi:10.1109/ICSE48619.2023.00200.

[64] eSafety Commissioner, Safety by Design, https://web.archive.org/web/20220308081249/https://www.esafety.gov.au/industry/safety-by-design (n.d.).

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

[66] T. Koscik, Identifying Abuse Vectors, https://web.archive.org/web/20220818200307/https://spinecone.gitbooks.io/identifying-abuse-vectors/content/ (2018).

[67] Journal of Online Trust and Safety, The Second Issue, Journal of Online Trust and Safety 1 (2) (2022).

[68] International Standards Organization, ISO 31000:2018(en), Risk management — Guidelines, https://www.iso.org/obp/ui/#iso:std:iso:31000 (2018).

[69] C. Johnson, L. Badger, D. Waltermire, J. Snyder, C. Skorupka, Guide to cyber threat information sharing, https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-150.pdf (2016).

[70] A. J. Ko, P. K. Chilana, Design, discussion, and dissent in open bug reports, in: Proceedings of the 2011 iConference, iConference '11, Association for Computing Machinery, New York, NY, USA, 2011, pp. 106–113. doi:10.1145/1940761.1940776.

[71] M. L. McHugh, Interrater reliability: The kappa statistic, Biochemia Medica 22 (3) (2012) 276–282.

[72] J. R. Landis, G. G. Koch, The measurement of observer agreement for categorical data, biometrics (1977) 159–174.

[73] Trust and Safety Foundation Project, Case Studies, http://web.archive.org/web/20230101182825/https://trustandsafetyfoundation.org/case-studies/ (n.d.).

[74] Meta, Transparency reports | Transparency Center, https://web.archive.org/web/20220308081251/https://transparency.fb.com/data/ (n.d.).

[75] Snapchat, Snapchat Transparency Report | Snapchat Transparency, https://web.archive.org/web/20220308081253/https://values.snap.com/privacy/transpa (n.d.).

[76] Discord, Transparency Reports, https://web.archive.org/web/20220308081258/https://discord.com/tags/transparency-reports (n.d.).

[77] D. Anandayuvaraj, M. Campbell, A. Tewari, J. C. Davis, Fail: Analyzing software failures from the news using llms, in: [ASE'24] Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering, 2024.

[78] TikTok, Guardian's Guide, https://web.archive.org/web/20230413213943/https://www.tiktok.com/safety/en/guardians-guide/ (Mar. 2021).

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

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

[81] C. Newton, The child safety problem on platforms is worse than we knew, https://web.archive.org/web/20230330144813/https://www.theverge.com/2021/5 safety-platforms-thorn-report-snap-facebook-youtube-tiktok (May 2021).

[82] T. Dowd, This Dangerous TikTok Challenge Just Killed a 12-Year-Old, https://web.archive.org/web/20230124210038/https://www.vice.com/en/article/pk blackout-challenge-kill-children (Jul. 2021).

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

[84] Information Commissioner's Office, ICO fines TikTok £12.7 million for misusing children's data, https://ico.org.uk/about-the-ico/media-centre/news-and-blogs/2023/04/ico-fines-tiktok-127-million-for-misusing-children-s-data/ (Apr. 2023).

[85] TikTok, TikTok News and Top Stories | TikTok Newsroom, https://web.archive.org/web/20230414084706/https://newsroom.tiktok.com/en-us/ (Aug. 2018).

[86] TikTok, TikTok's Top 10 Tips for Parents, https://web.archive.org/web/20221006001304/https://newsroom.tiktok.com/en-us/tiktoks-top-10-tips-for-parents (Oct. 2019).

[87] TikTok, TikTok for Younger Users, https://web.archive.org/web/20230322110612/https://newsroom.tiktok.com/en-us/tiktok-for-younger-users (Dec. 2019).

[88] J. Collins, TikTok introduces Family Pairing, https://web.archive.org/web/20230412022455/https://newsroom.tiktok.com/en-us/tiktok-introduces-family-pairing (Apr. 2020).

[89] T. Elizabeth, Our work to keep TikTok a place for people 13 and over, 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 (May 2021).

[90] T. Elizabeth, Our work to design an age-appropriate experience on TikTok, https://web.archive.org/web/20230130014809/https://newsroom.tiktok.com/en us/our-work-to-design-an-age-appropriate-experience-on-tiktok/ (May 2021).

[91] A. Evans, Furthering our safety and privacy commitments for teens on TikTok, https://web.archive.org/web/20221018153406/https://newsroom.tiktok.com au/furthering-our-safety-and-privacy-commitments-for-teens-tiktok (Aug. 2021).

[92] A. Evans, New Family Pairing resources offer digital safety advice from teens, https://web.archive.org/web/20230311000442/https://newsroom.tiktok.com us/new-family-pairing-resources-offer-digital-safety-advice-from-teens (Sep. 2021).

[93] TikTok, Effective in-app reporting for a healthy community, https://newsroom.tiktok.com/en-us/effective-in-app-reporting-for-a-healthy-community (Aug. 2019).

[94] TikTok, Protecting teens online, https://www.tiktok.com/transparency/en-gb/protecting-teens/ (Aug. 2021).

[95] TikTok, Protecting teens online, https://support.tiktok.com/en/using-tiktok/messaging-and-notifications/comments (Aug. 2021).

[96] Strengthening privacy and safety for youth on tiktok, https://newsroom.tiktok.com/en-us/strengthening-privacy-and-safety-for-youth (Aug. 2021).

[97] TikTok, Age-restricted content on tiktok live, https://www.bbc.com/news/technology-64813981 (Aug. 2021).

[98] Choosing between a private or public account, https://support.tiktok.com/en/account-and-privacy/account-privacy-settings/making-your-account-public-or-private (Aug. 2021).

[99] BBC, Tiktok sets 60-minute daily screen time limit for under-18s, https://www.bbc.com/news/technology-64813981 (Aug. 2021).

[100] Removing followers, https://support.tiktok.com/en/using-tiktok/followers-and-following/removing-followers (Aug. 2021).

[101] Blocking users, https://support.tiktok.com/en/using-tiktok/followers-and-following/blocking-the-users (Aug. 2021).

[102] D. Anandayuvaraj, 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.

[103] A. R. Ibrahimzada, Y. Varli, D. Tekinoglu, R. Jabbarvand, Perfect is the enemy of test oracle, in: A. Roychoudhury, C. Cadar, M. Kim (Eds.), 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, ACM, 2022, pp. 70–81. doi:10.1145/3540250.3549086. URL https://doi.org/10.1145/3540250.3549086

[104] R. Angell, B. Johnson, Y. Brun, 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.

[105] G. Salvendy (Ed.), Handbook of Human Factors and Ergonomics, 4th Edition, Wiley, Hoboken, NJ, 2012.

[106] J. C. Bastien, Usability testing: A review of some methodological and technical aspects of the method, International Journal of Medical Informatics 79 (4) (2010) e18–e23. doi:10.1016/j.ijmedinf.2008.12.004.

[107] Social Web Working Group, ActivityPub, https://w3c.github.io/activitypub/#security-considerations (Jan. 2018).

[108] J. H. Saltzer, D. P. Reed, D. D. Clark, End-to-end arguments in system design, ACM Transactions on Computer Systems (TOCS) 2 (4) (1984) 277–288.

[109] B. Kostova, S. Gürses, C. Troncoso, Privacy Engineering Meets Software Engineering. On the Challenges of Engineering Privacy By Design, arXiv:2007.08613 [cs] (Jul. 2020). arXiv:2007.08613.

[110] V. Casola, A. De Benedictis, M. Rak, U. Villano, A novel Security-by-Design methodology: Modeling and assessing security by SLAs with a quantitative approach, Journal of Systems and Software 163 (2020) 110537. doi:10.1016/j.jss.2020.110537.

[111] I. Rubinstein, N. Good, Privacy by Design: A Counterfactual Analysis of Google and Facebook Privacy Incidents, SSRN Electronic Journal (2012). doi:10.2139/ssrn.2128146.

[112] E. Kalliamvakou, G. Gousios, K. Blincoe, L. Singer, D. M. German, D. Damian, The promises and perils of mining GitHub, in: Proceedings of the 11th Working Conference on Mining Software Repositories, MSR 2014, Association for Computing Machinery, New York, NY, USA, 2014, pp. 92–101. doi:10.1145/2597073.2597074.

[113] J. Aranda, 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, 2009, pp. 298–308. doi:10.1109/ICSE.2009.5070530.

[114] W. Wang, D. Arya, N. Novielli, J. Cheng, 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, Association for Computing Machinery, New York, NY, USA, 2020, pp. 1–14.

[115] A. S. Mashiyat, M. Famelis, R. Salay, 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, RSSE 2014, Association for Computing Machinery, New York, NY, USA, 2014, pp. 1–5. doi:10.1145/2593822.2593823.

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

[117] S. E. Toulmin, The Uses of Argument, Cambridge University Press, 2003.

[118] R. R. Dirk, Argument Diagramming of Meeting Conversations (2005).