M.S. Thesis Defense:
An Empirical Study of Trust & Safety Engineering in Open-Source Social Media Platforms

Geoffrey Cramer
April 4, 2023

Committee
Dr. James C. Davis
Chair
Dr. Alex Quinn
Member
Dr. Alice Marwick
Member
A Third of TikTok’s U.S. Users May Be 14 or Under, Raising Safety Questions

The child safety problem on platforms is worrisome, a report finds.

This Dangerous TikTok Challenge Just Killed a 12-Year-Old

TikTok may push potentially harmful content to teens within minutes, study finds

By Samantha Murphy Kelly, CNN Business

Updated 4:21 PM EST, Thu December 15, 2022

July 23, 2021, 12:56pm
Outline

- Background
- Research Questions
- Methodology
- Results
- Discussion
- Threats to Validity
Outline

- Background
- Research Questions
- Methodology
- Results
- Discussion
- Threats to Validity
Social Media Platform (SMP): Definitions and Types

• Defined as "internet-based... persistent channel[s] of mass personal communication facilitating perceptions of interactions among users, deriving value primarily from user-generated content" (Carr & Hayes, 2014)

• Smith’s Honeycomb Model (Smith 2007)
Open-source Software SMPs

- Open-source Software (OSS) SMPs first appeared around 2010 and leveraged public sharing protocols.
- The most popular OSS SMP, Mastodon, appeared in 2016 and mimics Twitter.
Social Media Platform (SMP): Context Diagram

Product design, business needs, external governance, etc.

Software Engineer

Software Development Lifecycle

Policy

Moderation

(Automated) (Manual)

Social Media Platform

Features [62]
- Identity
- Presence
- Relationships
- Reputation
- Groups
- Conversations
- Sharing

User-generated Interactions

Filters
- Follow List
- Mute List
- Block List
- Keyword Lists
- Domain Lists
- Content Algorithms

Interaction oversight

Eve interacts with Bob

Interaction

Views interactions
Global surveys have found:

- Almost 3 in 5 people use social media (DataReportal, 2023)
- 48% of people experience hate & harassment (Thomas et al., 2021)
- Daily SMP users are 2x more likely to experience H&H than non-users (Thomas et al., 2021)

Trust & Safety (T&S) was invented to address these issues

- "the study of how people abuse the Internet to cause real human harm" (Cryst et al., 2021)
- With high rates of online abuse, T&S teams are still struggling

T&S Engineering emerged recently as a discipline to "design [software] with user safety in mind" (Galantino, 2019)
T&S in SMPs: A Risk Management View

• T&S is broad. I use a narrowed definition to scope the study
  • **User T&S in SMPs** is the study of how users harm other users on SMPs
  • **User T&S Engineering in SMPs** is a collection of software engineering methods that use T&S knowledge to reduce harmful user-to-user interaction

• T&S is uncertain. I use a risk management model to frame the study
  • **T&S Risk** is the potential loss that users face when harmed by other users
  • ISO 31000:2018 contains *risk assessment* and *risk treatment* steps
T&S in SMPs: Risk Assessment

- Taxonomy of hate & harassment (Thomas et al., 2021)
  - Toxic content
  - Content leakage
  - Overloading
  - False reporting
  - Impersonation
  - Surveillance
  - Lockout & control
T&S in SMPs: Risk Treatment

Product design, business needs, external governance, etc.

Design Treatment

Moderation Treatment

Features [62]
- Identity
- Presence
- Relationships
- Sharing
- Reputation

User-generated Interactions

Filters
- Follow List
- Mute List
- Block List
- Keyword Lists
- Domain Lists
- Content Algorithms

Social Media Platform

Policy

Policies

Moderation

(Automated) (Manual)

Interaction oversight

Views interactions
Design Treatment

- Interface-specific solutions
  - Socially-aware content access control (Misra & Such, 2016)
  - Forums designed with captchas (Seering et al., 2019)

- Encompassing solutions
  - Privacy & Security By Design (Cavoukian & Dixon, 2013)
  - Safety By Design (eSafety Commissioner, n.d.)
  - Abuse vector treatment patterns (Koscik, 2018)
Design Treatment – T&S By Design

Theory

Safety By Design [57]
- Harmful content detection
- Privacy & security by design
- Communicate social contracts
- Convey service guidelines
- Mitigate feature risk factors
- Leverage technical features
- Provide users with safety tools
- Robust content reporting

Practice

Koscik [18]
- Add moderation
- Require consent
- Remove data
- Interaction intervention
- Reduce visibility
- Reduce interaction
- Remove feature
Summary and Unknowns

• Prior academic work has
  • Taxonomized threats
  • Investigated context-specific solutions
• Prior grey literature has
  • Provided platform governance principles
  • Listed potential treatments to mitigate abuse
• No empirical work on how T&S Engineering is practiced
Outline

- Background
- Research Questions
- Methodology
- Results
- Discussion
- Threats to Validity
Research Questions

Risk Assessment

**RQ1** In what contexts do T&S engineering problems arise?

**RQ2** What risks are identified in T&S engineering discussions?

Risk Treatment

**RQ3** What options are proposed in T&S engineering discussions? How are they selected?

Risk-based Decision Model based on ISO 31000:2018
Outline

- Background
- Research Questions
- Methodology
- Results
- Discussion
- Threats to Validity
Methodology – Dataset Selection

• Goal: collect & analyze empirical data to characterize current T&S engineering processes

• OSS SMPs maintain issue tracking systems to track problems

• Accessible, traceable dataset of engineering decision processes

• Can filter issues to extract those related to T&S
Methodology – T&S Issue Example

Proposal: “Before you interact” modal (Open)

User A: before you reply modal can be shown before replying to non-mutual posts would remind users to follow certain etiquette in certain situations [similar] to the rules/guidelines shown when joining a Twitch stream
What might be interesting is to allow users to set their own prompts

User B: many people are using limited bio space to do so but are not easy to access

User C: If the message can be set by the user being replied to, it’s an avenue for abuse

User A: [can] show this in the admin view [so it is] moderable

User C: annoying types could use a client that does not implement this feature

User A: Sure, so user-defined rules need api/federation modification
Repository Selection

<table>
<thead>
<tr>
<th>Project</th>
<th>Category</th>
<th>Accounts</th>
<th>Issues</th>
<th>Stars</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mastodon</td>
<td>Microblogging</td>
<td>7,833,218</td>
<td>8,892</td>
<td>39.7K</td>
</tr>
<tr>
<td>Diaspora</td>
<td>Social networking</td>
<td>740,409</td>
<td>4,719</td>
<td>13.2K</td>
</tr>
<tr>
<td>PeerTube</td>
<td>Video sharing</td>
<td>288,964</td>
<td>4,386</td>
<td>11.4K</td>
</tr>
<tr>
<td>pixelfed</td>
<td>Photo sharing</td>
<td>150,326</td>
<td>1,702</td>
<td>4.5K</td>
</tr>
<tr>
<td>Pleroma</td>
<td>Microblogging</td>
<td>127,861</td>
<td>2,983</td>
<td>123</td>
</tr>
<tr>
<td>BirdsiteLive</td>
<td>Microblogging</td>
<td>101,188</td>
<td>91</td>
<td>398</td>
</tr>
</tbody>
</table>

OSS SMP projects with over 100K accounts. With number of GitHub Issues and Stars. Data in this table pulled on January 26, 2023.
Repository Selection → Issue Selection → Issue Analysis

- Discussion Modeling
  - Issue Metadata (RQ1)
  - Risk Statements (RQ2)
  - Treatment + Rationale Statements (RQ3)

→ T&S Engineering Process
Issue Selection

Trust & Safety Journal Keywords

Sanitization

Baseline Keywords (N=12)

Sample Issues (N=100)

Add Keywords if Recall < 90%

Tailored Keywords

Filtered Issues (Diaspora=316, Mastodon=431)

Randomly Order Dataset

Process until N=60

Baseline Keywords

Keyword Tailoring

Issue Sampling
### Issue Selection – Keyword Results

<table>
<thead>
<tr>
<th>Project</th>
<th>Keywords</th>
<th>Prec., Rec.</th>
<th>Filtered Issues</th>
<th>Analysis %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mastodon</td>
<td>17</td>
<td>50%, 100%</td>
<td>431</td>
<td>26%</td>
</tr>
<tr>
<td>Diaspora</td>
<td>15</td>
<td>27%, 100%</td>
<td>316</td>
<td>73%</td>
</tr>
</tbody>
</table>

SMP filtering results
Repository Selection → Issue Selection

Issue Analysis → Discussion Modeling

- Issue Metadata (RQ1)
- Risk Statements (RQ2)
- Treatment + Rationale Statements (RQ3)

T&S Engineering Process
### Issue Analysis – Codebooks

#### Codebook for issue type

<table>
<thead>
<tr>
<th>Label</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bug</td>
<td>A mistake in implementation that deviates from the original design intent.</td>
<td>“Can’t suspend users with + sign in their email address” (Mastodon #10576)</td>
</tr>
<tr>
<td>Feature request</td>
<td>A proposal for a new addition or modification to the system.</td>
<td>“Instance Greylisting” (Mastodon #4296)</td>
</tr>
</tbody>
</table>

#### Codebook for issue result

<table>
<thead>
<tr>
<th>Label</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open</td>
<td>Issue is still in the “open” state and is unresolved.</td>
</tr>
<tr>
<td>No action</td>
<td>Issue is closed but not change was made to the codebase.</td>
</tr>
<tr>
<td>Merged</td>
<td>Issue is closed with some change to the codebase.</td>
</tr>
</tbody>
</table>
## Issue Analysis – Codebooks (cont.)

<table>
<thead>
<tr>
<th>Label</th>
<th>Description</th>
<th>Criteria</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Risk</td>
<td>Claim of potential loss that users face when harmed by other users.</td>
<td>Specifically mentions a type of online abuse (e.g. harassment), a scenario that could lead to online abuse, or a weakness that leaves users open to abuse. Reiterated items are not re-coded.</td>
<td>“On Twitter, DMs became a terrible spam vector” (Mastodon #90)</td>
</tr>
<tr>
<td>Option</td>
<td>Proposal to progress the issue towards closure.</td>
<td>Implementation details or UI design are not re-coded.</td>
<td>“We could let the user decide if he wants to lock it down or not.” (Diaspora #798)</td>
</tr>
<tr>
<td>Chosen</td>
<td>An option that is selected by engineers.</td>
<td>If maintainers choose the associated option and close the issue, this code should be filled in.</td>
<td>“You can’t pin [content] from other accounts and you won’t be able to because it’s open to various forms of abuse” (Mastodon #5182)</td>
</tr>
<tr>
<td>Treatment selection rationale</td>
<td>Reason to select an option.</td>
<td>Specifies why a particular option should be selected and acted upon. Only coded for options that are marked as chosen.</td>
<td>“With the surge of new users …more people ought to be reviewing [content]” (Mastodon #811)</td>
</tr>
</tbody>
</table>
### Discussion Modeling Example 1

#### Appeals Function (Closed)

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

**Add ability to change a post scope after its publication (Closed)**

<table>
<thead>
<tr>
<th>ID</th>
<th>User</th>
<th>Comment</th>
<th>Option</th>
<th>Risk</th>
<th>Rationale</th>
<th>Chosen</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>“Add ability to change a post scope after it’s publication”</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>B</td>
<td>“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.”</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>3</td>
<td>B</td>
<td>“I was thinking of maybe allow to change visibility only if the post has no comment.”</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study Data</td>
<td>Related Work</td>
<td>Taxonomy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------------------</td>
<td>---------------------------------------------------</td>
<td>-----------------------------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>General issue topics</td>
<td>Honeycomb model (Smith, 2007)</td>
<td>SMP feature taxonomy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Risk sentences</td>
<td>Hate &amp; harassment taxonomy (Thomas et al., 2021)</td>
<td>T&amp;S risk taxonomy T&amp;S threat actor taxonomy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Option sentences</td>
<td>Abuse vector treatment taxonomy (Koscik, 2018)</td>
<td>T&amp;S Engineering pattern taxonomy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rationale sentences</td>
<td>Software quality rationale taxonomy (Ko &amp; Chilana, 2011)</td>
<td>T&amp;S treatment selection rationale taxonomy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Issue Analysis – Inter-rater agreement

<table>
<thead>
<tr>
<th>Study Data</th>
<th>Agreement method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentence codings</td>
<td><em>Risk</em> $\kappa = 0.89$, <em>Option</em> $\kappa = 1.0$, <em>Rationale</em> $\kappa = 1.0$, <em>Chosen</em> $\kappa = 1.0$</td>
</tr>
<tr>
<td>SMP feature taxonomy</td>
<td>Did not pursue inter-rater agreement.</td>
</tr>
<tr>
<td>T&amp;S Engineering pattern taxonomy</td>
<td>Random sample of 12/119. $\kappa = 0.73$</td>
</tr>
<tr>
<td>Rationale taxonomy</td>
<td>Random sample of 10%. $\kappa = 0.81$</td>
</tr>
<tr>
<td>Risk taxonomy</td>
<td>Collaborated on all categorizations.</td>
</tr>
</tbody>
</table>
Outline

• Background
• Research Questions
• Methodology
• Results
• Discussion
• Threats to Validity
RQ1: In what contexts do T&S engineering problems arise?
T&S issue curve is distinct
<table>
<thead>
<tr>
<th>Feature</th>
<th>Element(s) [20]</th>
<th>Diaspora</th>
<th>Mastodon</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>moderation</td>
<td>Infrastructure</td>
<td>4</td>
<td>8</td>
<td>12</td>
</tr>
<tr>
<td>content sharing</td>
<td>Sharing</td>
<td>9</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>user registration</td>
<td>Identity</td>
<td>6</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>private messaging</td>
<td>Conversations,</td>
<td>3</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Groups</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>content tagging</td>
<td>Sharing</td>
<td>3</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>user relationships</td>
<td>Relationships</td>
<td>4</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>content filters</td>
<td>Sharing</td>
<td>0</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>user filters</td>
<td>Presence,</td>
<td>0</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Relationships</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>instance filters</td>
<td>Groups</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>content metadata</td>
<td>Sharing</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>user profile</td>
<td>Identity</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
RQ1: Key Findings

• Both projects see T&S issue frequency rise 1-2 years after project creation

• Distinct feature concerns between two platforms

• 92% of T&S issues were feature requests instead of bugs

• 13 out of 60 T&S issues referenced other SMPs
RQ2: What risks are identified in T&S engineering discussions?
Threat Actor Analysis

• Each risk sentence was associated with a threat actor.
  • **User**: “captcha will remind user that this is serious and will avoid spamming.” (Diaspora #4711)
  • **Moderator**: “Moderators [can] access private [content]” (Mastodon #6986)
  • **Bot**: “current one is very bad at preventing bot registrations” (Diaspora #8342)
  • **External Actor**: “risk of a hostile instance harvesting private messages” (Mastodon #4296)

• Over 50% of risk statements identify user as a threat actor.

• 20% of risk statements identify moderator as a threat actor.
### T&S Risk Taxonomy

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

- False reporting
- Surveillance
- Lockout & Control
- Overmod.*
- Undermod.*
- Imp./Faulty Accts.*
- Other
- Content Leakage
- Toxic Content
- Overloading

Legend:
- Content Sharing
- Moderation
RQ2: Key Findings

- Extended risk taxonomy: under- and over-moderation, impersonation / faulty accounts
- Mastodon is primarily concerned with toxic content. Diaspora with content leakage
- Risk landscapes can vary based on each feature
- Moderation issues are difficult to resolve due to a diverse risk landscape
RQ3: What options are proposed in T&S engineering discussions? How are they selected?
<table>
<thead>
<tr>
<th>Pattern</th>
<th>Description</th>
<th>Proposed</th>
<th>Chosen</th>
</tr>
</thead>
<tbody>
<tr>
<td>Add moderation</td>
<td>Add or improve moderation tools</td>
<td>20</td>
<td>7</td>
</tr>
<tr>
<td>Require consent</td>
<td>Ask for approval from involved stakeholders</td>
<td>15</td>
<td>4</td>
</tr>
<tr>
<td>Improve filters</td>
<td>Allow users to better control the content they see</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>Reduce visibility</td>
<td>Limit when a feature can be used</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>Improve registration</td>
<td>Bolster user trustworthiness checks</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>Reduce audience</td>
<td>Limit exposure of content</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>Interaction intervention</td>
<td>Intervene before users contact others</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Moderation transparency</td>
<td>Increase clarity of moderation decisions</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Interaction transparency</td>
<td>Clarity of events that occurred between users</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Remove data</td>
<td>Remove unnecessary data from platform</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Reduce interaction</td>
<td>Limit how a feature can be used</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Remove feature</td>
<td>Take out feature</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
T&S Engineering Pattern Diagram

Old context diagram. Will now overlay discovered patterns...
T&S Engineering Pattern Diagram

Legend
- fully automated
- relies on Alice
- relies on Bob
- relies on Moderator
- Pattern (chosen/proposed)
- New Pattern

Proactive
- Interaction Intervention (1/5)
- Features [20]
  - Identity
  - Presence
  - Relationships
  - Reputation
  - Groups
  - Conversations
  - Sharing
- Reduce Interaction (0/2)
- Remove Feature (0/0)

Reactive
- Interaction Transparency (0/4)
- Reduce Audience (1/7)
- Moderation Transparency (0/5)
  - Moderation (Automated)
  - Add Moderation (8/35)
- Filters
  - Followed Users
  - Mute List
  - Block List
  - Keyword Lists
  - Domain Lists
  - Content Algorithms
- Improve Filters (4/16)

Eve interacts with
- Improve Registration (3/8)
- Reduce Visibility (1/12)
- Require Consent (4/21)
- Remove Data (0/4)

Bob views interactions
T&S Engineering Pattern Diagram

Legend
- fully automated
- relies on Alice
- relies on Bob
- relies on Moderator
- Pattern (chosen/proposed)
- New Pattern

Proactive
- Interaction Intervention (1/5)
- Improve Registration (3/8)
- Reduce Visibility (1/12)
- Require Consent (4/21)
- Reduce Interaction (0/2)
- Remove Feature (0/0)
- Remove Data (0/4)

Features [20]
- Identity
- Presence
- Relationships
- Reputation
- Groups
- Conversations
- Sharing

Reactive
- Interaction Transparency (0/4)
- Reduce Audience (1/7)
- Moderation Transparency (0/5)
- Moderation (Automated)
- (Manual)
- Add Moderation (8/35)

Filters
- Followed Users
- Mute List
- Block List
- Keyword Lists
- Domain Lists
- Content Algorithms

Eve interacts with Bob

Bob views interactions

interaction oversight

user-generated interactions
T&S Engineering Pattern Diagram

Legend
- fully automated
- relies on Alice
- relies on Bob
- relies on Moderator
- Pattern (chosen/proposed)

Proactive
- Interaction Intervention (1/5)
  - Reduce Interaction (0/2)
  - Remove Feature (0/0)
- Features [20]
  - Identity
  - Presence
  - Relationships
  - Reputation
  - Groups
  - Conversations
  - Sharing
- Remove Data (0/4)

Reactive
- Interaction Transparency (0/4)
- Reduce Audience (1/7)
- Moderation Transparency (0/5)
  - Add Moderation (8/35)

Filters
- Followed Users
- Mute List
- Block List
- Keyword Lists
- Domain Lists
- Content Algorithms

Interaction oversight
- views interactions
A Third of TikTok’s U.S. Users May Be 14 or Under, Raising Safety Questions

By Samantha Murphy Kelly, CNN Business

Updated 4:21 PM EST, Thu December 15, 2022

July 23, 2021, 12:56pm

The child safety problem on platforms is worse than we thought, a new study finds.

VICE News

This Dangerous TikTok Challenge Just Killed a 12-Year-Old

TikTok may push potentially harmful content to teens within minutes, study finds.

By Samantha Murphy Kelly, CNN Business

Updated 4:21 PM EST, Thu December 15, 2022

July 23, 2021, 12:56pm
TikTok introduces Family Pairing

By Jeff Collins, Trust & Safety, San Francisco Bay Hub

- Parental controls -> require consent
- Daily time limit -> reduce visibility
- Limit inappropriate content -> improve filters
- Restrict who can send DMs -> require consent
- Turn off DMs completely -> reduce visibility
Our work to keep TikTok a place for people 13 and over

- Prevent underage users from signing up -> improve verification
- Remove underage accounts -> add moderation
- Bring transparency to our actions -> moderation transparency
- Age-appropriate environment for new young users
  - Remove access to LIVE -> reduce visibility
  - Remove access to DMs -> reduce visibility
## Rationale Taxonomy

<table>
<thead>
<tr>
<th>Result</th>
<th>Rationale</th>
<th>Description</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>MERGED</td>
<td>Safety</td>
<td>Protects user from T&amp;S risks</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>Efficiency</td>
<td>Easy completion of tasks</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Mod. efficiency</td>
<td>Easy completion of admin/mod tasks</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>User efficiency</td>
<td>Easy completion of SMP user tasks</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Feasibility</td>
<td>Ease of implementation</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Flexibility</td>
<td>Handles a variety of use cases</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Clarity</td>
<td>Provides clear experience to users</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Security</td>
<td>Prevents unwanted data access</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Annoyance</td>
<td>Removes hindrance to user activity</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td><strong>Infeasibility</strong></td>
<td>Difficulty of implementation</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td><strong>Internal Infeasibility</strong></td>
<td>Difficulty due to internal factors</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td><strong>External Infeasibility</strong></td>
<td>Difficulty due to external factors</td>
<td>6</td>
</tr>
<tr>
<td>NO ACTION</td>
<td><strong>Unsafety</strong></td>
<td>Adverse effect to user T&amp;S</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Insecure</td>
<td>Susceptible to unwanted data access</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Inconsistency</td>
<td>Conflicts with design or user expectations</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td><strong>Uncertainty</strong></td>
<td>Unclear design or T&amp;S environment</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Annoyance</td>
<td>Unnecessary hindrance to user activity</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Unclarity</td>
<td>Convoluted user experience</td>
<td>1</td>
</tr>
</tbody>
</table>
Diaspora is more passive. 40% of issues are open.
RQ3: Key Findings

- Most commonly proposed and chosen pattern is *add moderation*
- Reactive patterns are chosen more often
- New patterns are primarily reactive
- 38% of identified T&S issues remain open
- T&S issue resolution is slow
Outline

• Background
• Research Questions
• Methodology
• Results
• Discussion
• Threats to Validity
Recommendations for OSS SMPs

- Communicate existing risks
- Document risks and treatments
- Explore proactive solutions
What other patterns could be used?

- Interaction intervention
  - Show timely messages to younger users
- Reduce visibility
  - Don’t allow unverified accounts to interact with young users
- Reduce audience + interaction transparency
  - Limit exposure of younger users’ content. Tell child + parent

TikTok introduces Pairing

By Jeff Collins, Trust & Safety, San Francisco Bay Hub

Our work to keep TikTok a place for people 13 and over
Future Work

- T&S Engineering Pattern Catalog
- Improve T&S Testing
- Automated content moderation in OSS SMPs
- T&S improvements in federated protocols
- T&S By Design
Remember this from the background?

Safety By Design [57]

- Harmful content detection
- Privacy & security by design
- Communicate social contracts
- Convey service guidelines
- Mitigate feature risk factors
- Leverage technical features
- Provide users with safety tools
- Robust content reporting

Koscik [18]

- Add moderation
- Require consent
- Remove data
- Interaction intervention
- Reduce visibility
- Reduce interaction
- Remove feature
Future Work – T&S By Design

Safety By Design

- Harmful content detection
- Privacy & security by design
- Communicate social contracts
- Convey service guidelines
- Mitigate feature risk factors
- Leverage technical features
- Provide users with safety tools
- Robust content reporting

T&S Engineering Patterns

- Add moderation
- Require consent
- Remove data
- Interaction intervention
- Reduce visibility
- Reduce interaction
- Remove feature

- Improve filters
- Moderation transparency
- Interaction transparency
- Reduce audience
- Improve registration

Koschik [18]
Outline

- Background
- Research Questions
- Methodology
- Results
- Discussion
- Threats to Validity
Threats to Validity

- Internal validity
  - Qualitative study
  - Diaspora 3rd-party forum
- External validity
  - Generalizability to commercial SMPs
  - Small sample size.
- Construct validity
  - T&S is vague
References


Galantino, L. (Director). (2019, May 30). Trust & Safety Engineering @ GitHub. https://www.youtube.com/watch?v=UC3Y9rx1jFQ&t=190


Thank You!
Bonus Slides
Mastodon accounts over time []

Elon Musk buys Twitter
Issue Selection

Trust & Safety Journal Keywords

Sanitization

Baseline Keywords (N=12)

Sample Issues (N=100)

Add Keywords if Recall < 90%

Tailored Keywords

Filtered Issues (Diaspora=316, Mastodon=431)

Randomly Order Dataset

Process until N=60

Baseline Keywords

Keyword Tailoring

Issue Sampling
Issue Selection

Trust & Safety Journal Keywords

Sanitization

Baseline Keywords (N=12)

Sample Issues (N=100)

Add Keywords if Recall < 90%

Tailored Keywords

Filtered Issues (Diaspora=316, Mastodon=431)

Randomly Order Dataset

Process until N=60

Keyword Tailoring

Issue Sampling
Issue Selection

**Baseline Keywords**

- Trust & Safety Journal Keywords
- Sanitization
- Baseline Keywords (N=12)

**Keyword Tailoring**

- Sample Issues (N=100)
- Add Keywords if Recall < 90%
- Tailored Keywords

**Issue Sampling**

- Filtered Issues (Diaspora=316, Mastodon=431)
- Randomly Order Dataset
- Process until N=60
### Issue Selection – Keyword Results

<table>
<thead>
<tr>
<th>Project</th>
<th>Keywords</th>
<th>Prec., Rec.</th>
<th>Filtered Issues</th>
<th>Analysis %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mastodon</td>
<td>17</td>
<td>50%, 100%</td>
<td>431</td>
<td>26%</td>
</tr>
<tr>
<td>Diaspora</td>
<td>15</td>
<td>27%, 100%</td>
<td>316</td>
<td>73%</td>
</tr>
</tbody>
</table>

SMP filtering results
Issue Selection – Eligible Population

Keyword Baselining

N≥5 Comments

Keyword Tailoring

Issue Sampling
## Issue Analysis – Issue Data Mapping

<table>
<thead>
<tr>
<th>GitHub Data</th>
<th>Use in study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Issue Title</td>
<td>Used for discussion modeling (if needed).</td>
</tr>
<tr>
<td>Issue Status (open/closed)</td>
<td>Used for <em>issue result</em> (&quot;open&quot;, &quot;merged&quot;, &quot;no action&quot;)</td>
</tr>
<tr>
<td>Issue open date</td>
<td>Used for longitudinal analysis.</td>
</tr>
<tr>
<td>Issue closure date</td>
<td>Used for longitudinal analysis</td>
</tr>
<tr>
<td>Issue comments</td>
<td>Separated comments into sentences.</td>
</tr>
<tr>
<td></td>
<td>Used to determine <em>issue type</em> (&quot;bug&quot; or &quot;feature request&quot;).</td>
</tr>
<tr>
<td></td>
<td>Used to label SMP feature.</td>
</tr>
<tr>
<td></td>
<td>Coded relevant sentences as <em>risk, option, rationale, and chosen</em>.</td>
</tr>
<tr>
<td>Linked pull requests</td>
<td>Used to distinguish between &quot;merged&quot; and &quot;no action&quot; <em>issue result</em>.</td>
</tr>
<tr>
<td></td>
<td>Used to determine which <em>options</em> to mark as <em>chosen</em>.</td>
</tr>
</tbody>
</table>
Inter-rater Agreement – Kappa Calculation

\[ \kappa = \frac{p_o - p_e}{1 - p_e} \]

- \( p_o \) is the relative observed agreement among raters
- \( p_e \) is the hypothetical probability of chance

<table>
<thead>
<tr>
<th>first</th>
<th>second</th>
<th>correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>improve filters</td>
<td>improve filters</td>
<td>1</td>
</tr>
<tr>
<td>reduce visibility</td>
<td>reduce visibility</td>
<td>1</td>
</tr>
<tr>
<td>improve filters</td>
<td>improve filters</td>
<td>1</td>
</tr>
<tr>
<td>require consent</td>
<td>interaction transparency</td>
<td>0</td>
</tr>
<tr>
<td>require consent</td>
<td>require consent</td>
<td>1</td>
</tr>
<tr>
<td>require consent</td>
<td>reduce audience</td>
<td>0</td>
</tr>
<tr>
<td>add moderation</td>
<td>add moderation</td>
<td>1</td>
</tr>
<tr>
<td>add moderation</td>
<td>add moderation</td>
<td>1</td>
</tr>
<tr>
<td>reduce audience</td>
<td>reduce audience</td>
<td>1</td>
</tr>
<tr>
<td>add moderation</td>
<td>add moderation</td>
<td>1</td>
</tr>
<tr>
<td>improve filters</td>
<td>improve filters</td>
<td>1</td>
</tr>
<tr>
<td>improve registration</td>
<td>add moderation</td>
<td>0</td>
</tr>
</tbody>
</table>

\[ k = \frac{(H14-1/12)/(1-1/12)}{0.7272727} \]
## Issue Results

<table>
<thead>
<tr>
<th>result</th>
<th>count</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>merged</td>
<td>19</td>
<td>0.31666667</td>
</tr>
<tr>
<td>no action</td>
<td>18</td>
<td>0.3</td>
</tr>
<tr>
<td>open</td>
<td>23</td>
<td>0.38333333</td>
</tr>
</tbody>
</table>
## Features – Diaspora vs. Mastodon

<table>
<thead>
<tr>
<th>feature</th>
<th>diaspora</th>
<th>mastodon</th>
</tr>
</thead>
<tbody>
<tr>
<td>user registration</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>content sharing</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>content tagging</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>private messaging</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>user relationships</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>user profile</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>moderation</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>user filters</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>instance filters</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>content filters</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>content metadata</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>
# Features Over Time

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>user registration</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>content sharing</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>content tagging</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>private messaging</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>user relationships</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>user profile</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>moderation</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>user filters</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>instance filters</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>content filters</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>content metadata</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>Overloading</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Toxic Content</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>7</td>
<td>7</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Content Leakage</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Other</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Impersonation / Faulty Accounts*</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Undermoderation*</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>6</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Overmoderation*</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Lockout and Control</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Surveillance</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>False reporting</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>#VALUE!</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
## Risks – Diaspora vs. Mastodon

<table>
<thead>
<tr>
<th>risk</th>
<th>diaspora</th>
<th>mastodon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overloading</td>
<td>6</td>
<td>11</td>
</tr>
<tr>
<td>Toxic Content</td>
<td>5</td>
<td>22</td>
</tr>
<tr>
<td>Content Leakage</td>
<td>19</td>
<td>5</td>
</tr>
<tr>
<td>Other</td>
<td>5</td>
<td>11</td>
</tr>
<tr>
<td>Impersonation / Faulty Accounts*</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Undermoderation*</td>
<td>6</td>
<td>11</td>
</tr>
<tr>
<td>Overmoderation*</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Lockout and Control</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Surveillance</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>False reporting</td>
<td>6</td>
<td>6</td>
</tr>
</tbody>
</table>
Patterns – excluded options examples

• "private messaging would be helpful"
• "I would suggest more [profile] bio space"
• "When someone #tags their comments, they should be visible on a #tag page"
• "Another solution could be if the username isn't found on the [instance], we make a search to other[s]"
• "Or even better, a choice of email or captcha, so anyone who doesn't want to use email has an option, as does anyone who hates captchas."
## Patterns Over Time

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>require consent</td>
<td>0</td>
<td>6</td>
<td>2</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>improve filters</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>reduce interaction</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>reduce visibility</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>interaction transparency</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>improve registration</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>add moderation</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>5</td>
<td>14</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>remove data</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>5</td>
<td>14</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>interaction intervention</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>moderation transparency</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>reduce audience</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
**Patterns – Diaspora vs. Mastodon**

<table>
<thead>
<tr>
<th>pattern</th>
<th>diaspora</th>
<th>mastodon</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>require consent</td>
<td>13</td>
<td>8</td>
</tr>
<tr>
<td>improve filters</td>
<td>0</td>
<td>16</td>
</tr>
<tr>
<td>reduce interaction</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>reduce visibility</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>interaction transparency</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>improve registration</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>add moderation</td>
<td>15</td>
<td>20</td>
</tr>
<tr>
<td>remove data</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>interaction intervention</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>moderation transparency</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>reduce audience</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>
T&S Engineering Pattern Diagram

Legend
- fully automated
- relies on Alice
- relies on Bob
- relies on Moderator
- Pattern (chosen/proposed)

Proactive
- Interaction Intervention (1/5)
- Features [20]
  - Identity
  - Presence
  - Relationships
  - Reputation
  - Groups
  - Conversations
  - Sharing
- Reduce Interaction (0/2)
- Remove Feature (0/0)
- Remove Data (0/4)
- Improve Registration (3/8)
- Reduce Visibility (1/12)
- Require Consent (4/21)
- New Pattern

Reactive
- Interaction Transparency (0/4)
- Reduce Audience (1/7)
- Improves Filters (4/16)
- Improvement Oversight
- Moderation
  - (Automated)
  - (Manual)
- Add Moderation (8/35)
- Followed Users
- Mute List
- Block List
- Keyword Lists
- Domain Lists
- Content Algorithms
- user-generated interactions
  - views interactions

Eve

Bob
T&S Engineering Pattern Diagram

Legend
- fully automated
- relies on Alice
- relies on Bob
- relies on Moderator
- Pattern (chosen/proposed)
- New Pattern

Proactive
- Reduce Interaction (0/2)
- Remove Feature (0/0)
- Interaction Intervention (1/5)
- Improve Registration (3/8)
- Reduce Visibility (1/12)
- Require Consent (4/21)
- Remove Data (0/4)
- Features [20]
  - Identity
  - Presence
  - Relationships
  - Reputation
  - Groups
  - Conversations
  - Sharing

Reactive
- Interaction Transparency (0/4)
- Reduce Audience (1/7)
- Improve Filters (4/16)
- Moderation Transparency (0/5)
  - (Automated)
  - (Manual)
- Add Moderation (8/35)

Filters
- Followed Users
- Mute List
- Block List
- Keyword Lists
- Domain Lists
- Content Algorithms

Eve interacts with Bob
- user-generated interactions
- views interactions

Interaction oversight
- interaction oversight
T&S Engineering Pattern Diagram

Legend
- fully automated
- relies on Alice
- relies on Bob
- relies on Moderator
- Pattern (chosen/proposed)
- New Pattern

Proactive
- Interaction Intervention (1/5)
- Reduce Interaction (0/2)
- Remove Feature (0/0)
- Features [20]
  - Identity
  - Presence
  - Relationships
  - Reputation
  - Groups
  - Conversations
  - Sharing
  - Remove Data (0/4)

Reactive
- Interaction Transparency (0/4)
- Reduce Audience (1/7)
- Moderation Transparency (0/5)
- Moderation (Automated) (Manual)

Filters
- Followed Users
- Mute List
- Block List
- Keyword Lists
- Domain Lists
- Content Algorithms

Interaction oversight
- views interactions

Eve interacts with
- Improve Registration (3/8)
- Reduce Visibility (1/12)
- Require Consent (4/21)

Bob
T&S Engineering Pattern Diagram

Legend
- fully automated
- relies on Alice
- relies on Bob
- relies on Moderator
- Pattern (chosen/proposed)
- New Pattern

Proactive
- Improve Registration (3/8)
- Reduce Visibility (1/12)
- Require Consent (4/21)
- Interact on Intervention (1/5)
- Features [20]
  - Identity
  - Presence
  - Relationships
  - Reputation
  - Groups
  - Conversations
  - Sharing
- Reduce Interaction (0/2)
- Remove Feature (0/0)

Reactive
- Interaction Transparency (0/4)
- Reduce Audience (1/7)
- Features [20]
  - Followed Users
  - Mute List
  - Block List
  - Keyword Lists
  - Domain Lists
  - Content Algorithms

Filters
- Improve Filters (4/16)

Moderation
- (Automated)
- (Manual)
- Add Moderation (8/35)

Eve interacts with Bob

User-generated interactions

Views interactions

Interaction oversight

Interaction interventions
Data Availability

- Artifact location: [https://zenodo.org/record/7601293](https://zenodo.org/record/7601293)
- GitHub issue mining tool
- Research data
  - Baseline keywords
  - Keyword tailoring process
  - Issue sampling
  - Discussion modeling
  - Taxonomy development
  - Codebook
  - Inter-rater agreement