M.S. Thesis Defense:

An Empirical Study of Trust & Safety Engineering in Open-Source Social Media Platforms

Geoffrey Cramer

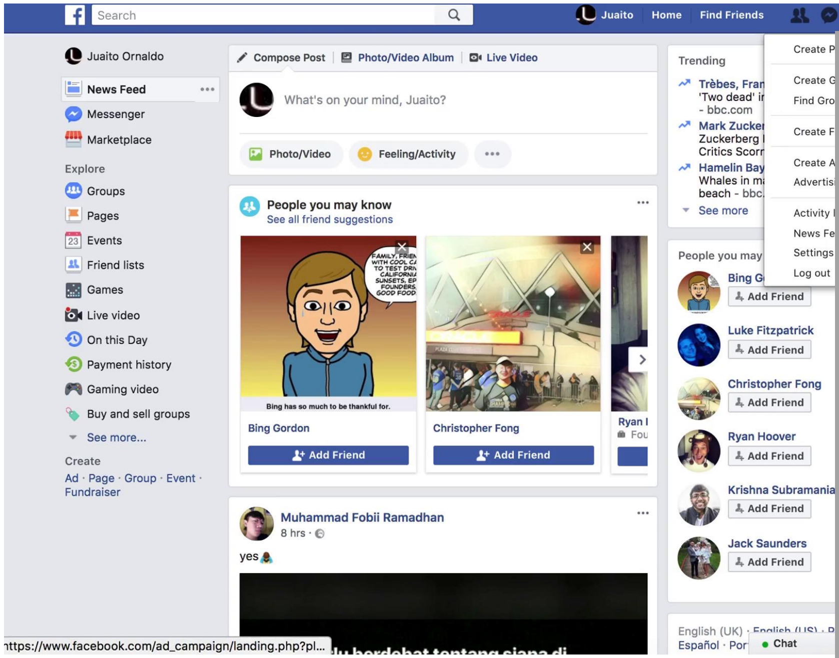
April 4, 2023

Committee

Dr. James C. Davis Chair	Dr. Alex Quinn Member	Dr. Alice Marwick Member

PURDUE UNIVERSITY

Elmore Family School of Electrical and Computer Engineering



Search

Juaito Home Find Friends

Juaito Ornaldo

News Feed

Messenger

Marketplace

Explore

Groups

Pages

Events

Friend lists

Games

Live video

On this Day

Payment history

Gaming video

Buy and sell groups

See more...

Create

Ad · Page · Group · Event · Fundraiser

Compose Post | Photo/Video Album | Live Video

What's on your mind, Juaito?

Photo/Video | Feeling/Activity

People you may know

See all friend suggestions

Bing has so much to be thankful for.

Bing Gordon — Add Friend

Christopher Fong — Add Friend

Ryan I

Muhammad Fobii Ramadhan

8 hrs ·

yes

Trending

Trèbes, Fran
'Two dead' i
- bbc.com

Mark Zucker
Zuckerberg
Critics Scorr

Hamelin Bay
Whales in m
beach - bbc.

See more

People you may

Bing G — Add Friend

Luke Fitzpatrick — Add Friend

Christopher Fong — Add Friend

Ryan Hoover — Add Friend

Krishna Subramania — Add Friend

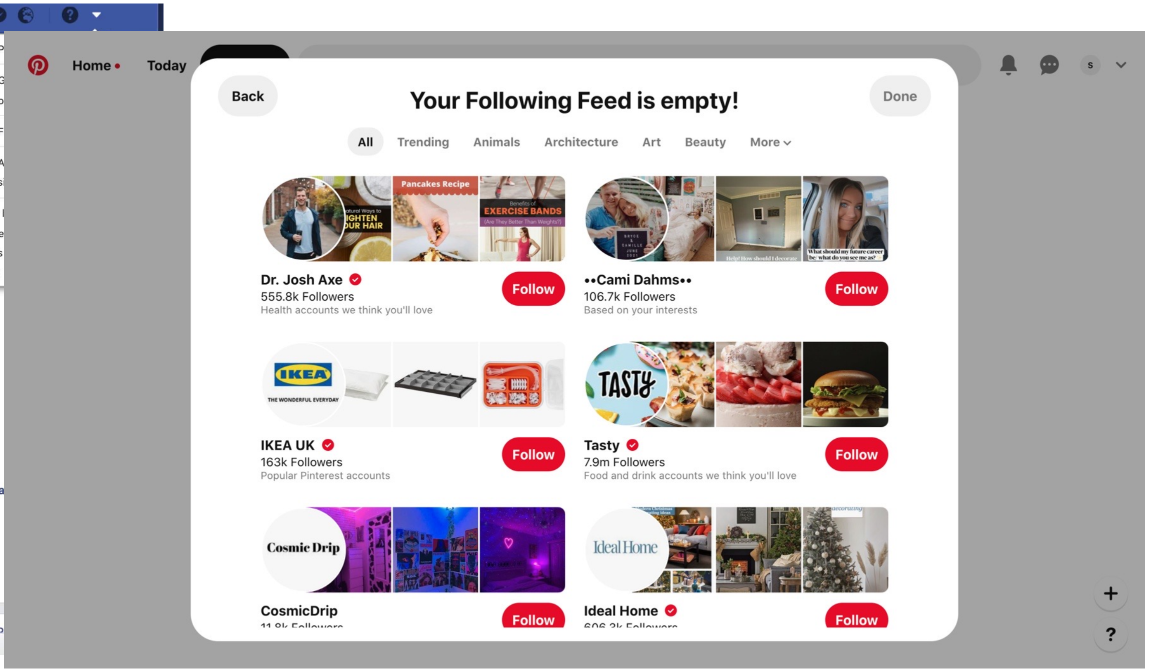
Jack Saunders — Add Friend

Create P
Create G
Find Gro
Create F
Create A
Advertis
Activity
News Fe
Settings
Log out

English (UK) · English (US) · P
Español · Por

Chat

https://www.facebook.com/ad_campaign/landing.php?pl...



Home • Today

Back

Your Following Feed is empty!

Done

All | Trending | Animals | Architecture | Art | Beauty | More

Dr. Josh Axe
555.8k Followers
Health accounts we think you'll love
Follow

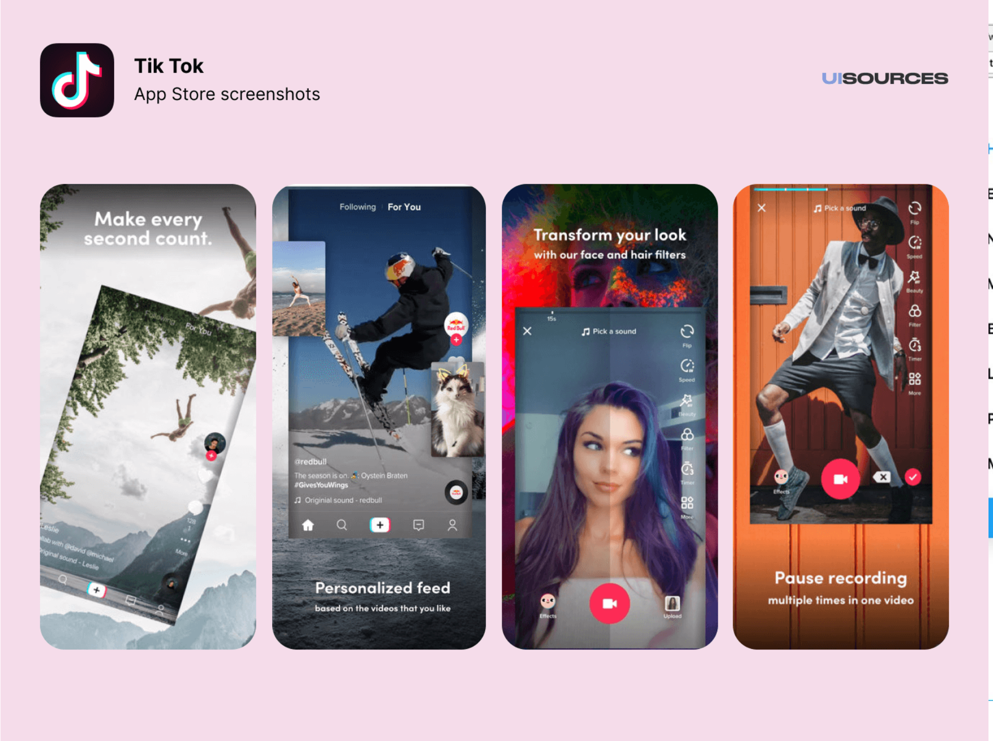
••Cami Dahms••
106.7k Followers
Based on your interests
Follow

IKEA UK
163k Followers
Popular Pinterest accounts
Follow

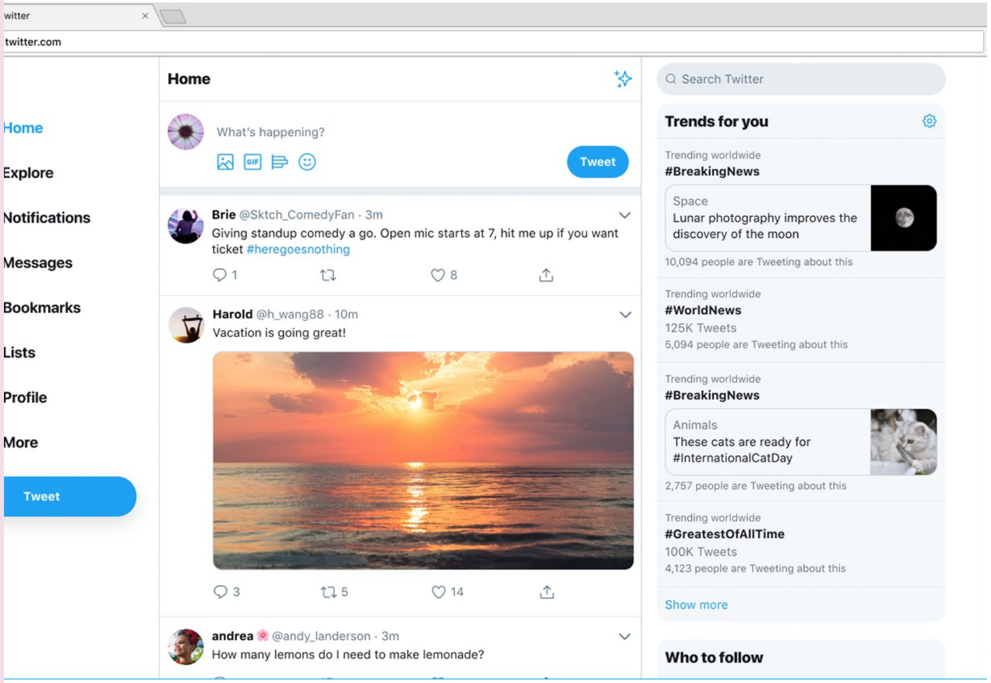
Tasty
7.9m Followers
Food and drink accounts we think you'll love
Follow

CosmicDrip
Follow

Ideal Home
Follow



Tik Tok
App Store screenshots



twitter.com

Home

Explore

Notifications

Messages

Bookmarks

Lists

Profile

More

Tweet

Home

What's happening?

Tweet

Brie @Sktch_ComedyFan · 3m
Giving standup comedy a go. Open mic starts at 7, hit me up if you want ticket #heregoesnothing
1 8

Harold @h_wang88 · 10m
Vacation is going great!
3 5 14

andrea @andy_landerson · 3m
How many lemons do I need to make lemonade?

Search Twitter

Trends for you

Trending worldwide
#BreakingNews
Space
Lunar photography improves the discovery of the moon
10,094 people are Tweeting about this

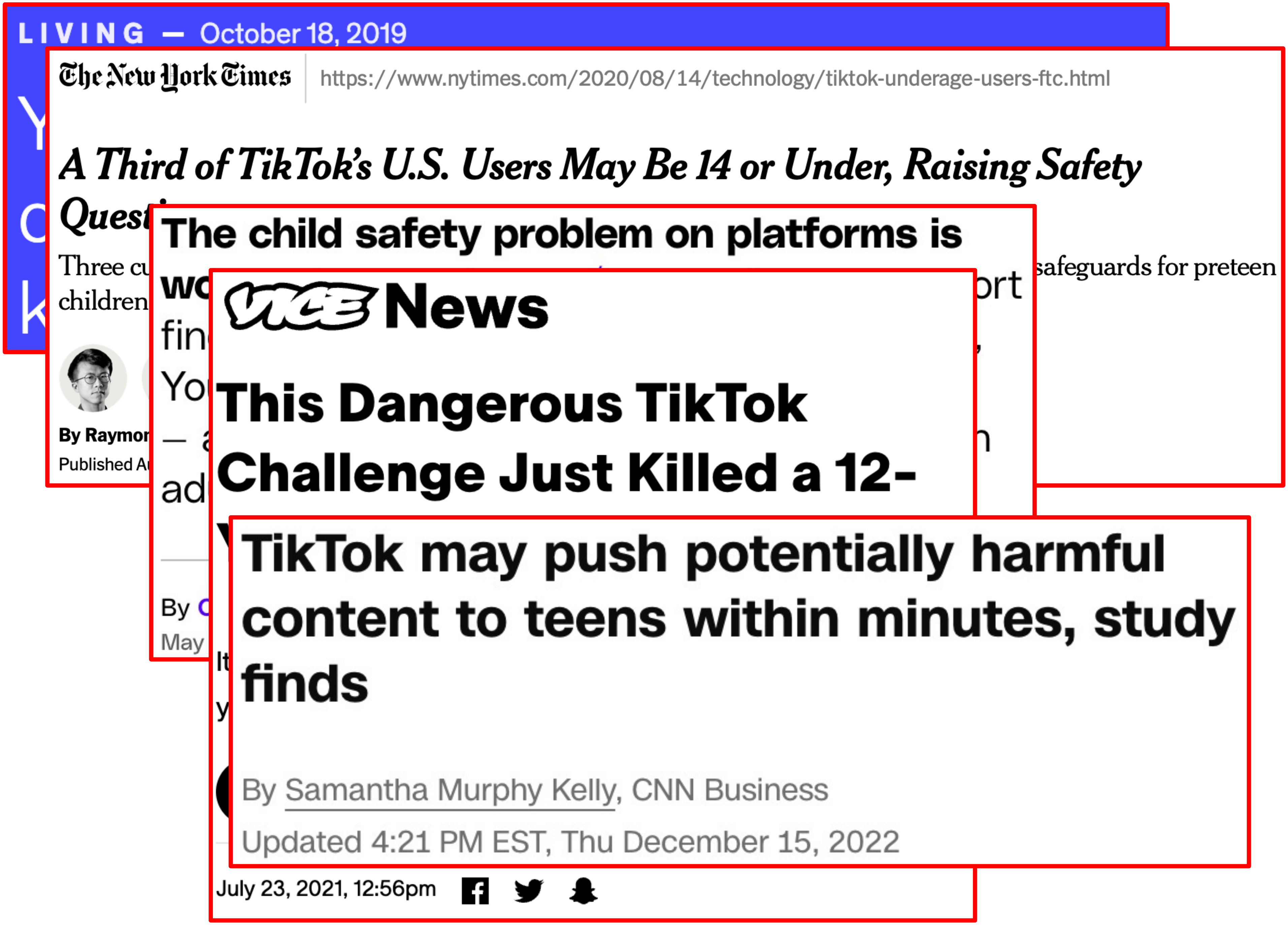
Trending worldwide
#WorldNews
125K Tweets
5,094 people are Tweeting about this

Trending worldwide
#BreakingNews
Animals
These cats are ready for #InternationalCatDay
2,757 people are Tweeting about this

Trending worldwide
#GreatestOfAllTime
100K Tweets
4,123 people are Tweeting about this

Show more

Who to follow



LIVING — October 18, 2019

Y
C
k

The New York Times https://www.nytimes.com/2020/08/14/technology/tiktok-underage-users-ftc.html

A Third of TikTok's U.S. Users May Be 14 or Under, Raising Safety Questi

Three cu safeguards for preteen children

By Raymon
Published A

The child safety problem on platforms is
wo ort
fin ,
Yo
— a n
ad

By C
May

VICE News

This Dangerous TikTok Challenge Just Killed a 12-
Y

It
y

July 23, 2021, 12:56pm

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

Outline

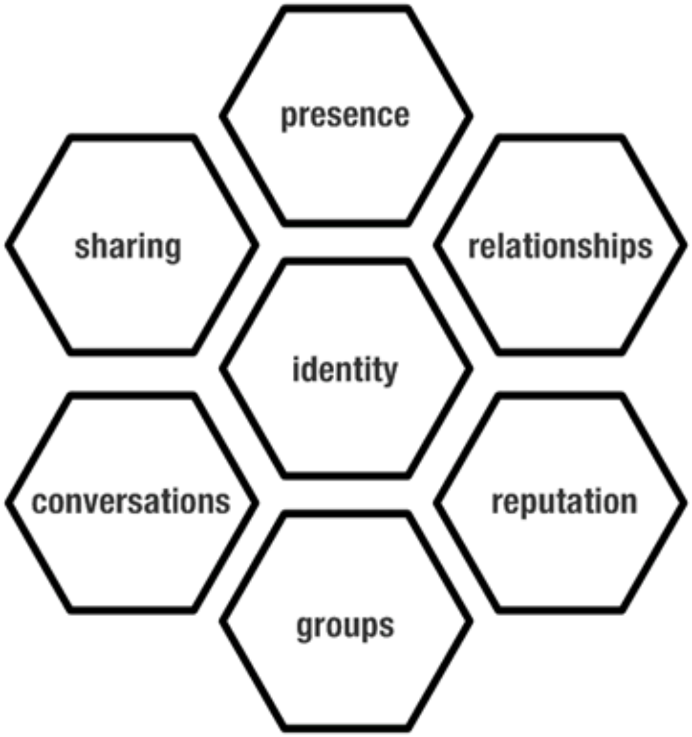
- Background
- Research Questions
- Methodology
- Results
- Discussion
- Threats to Validity

Outline

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

Social Media Platform (SMP): Definitions and Types

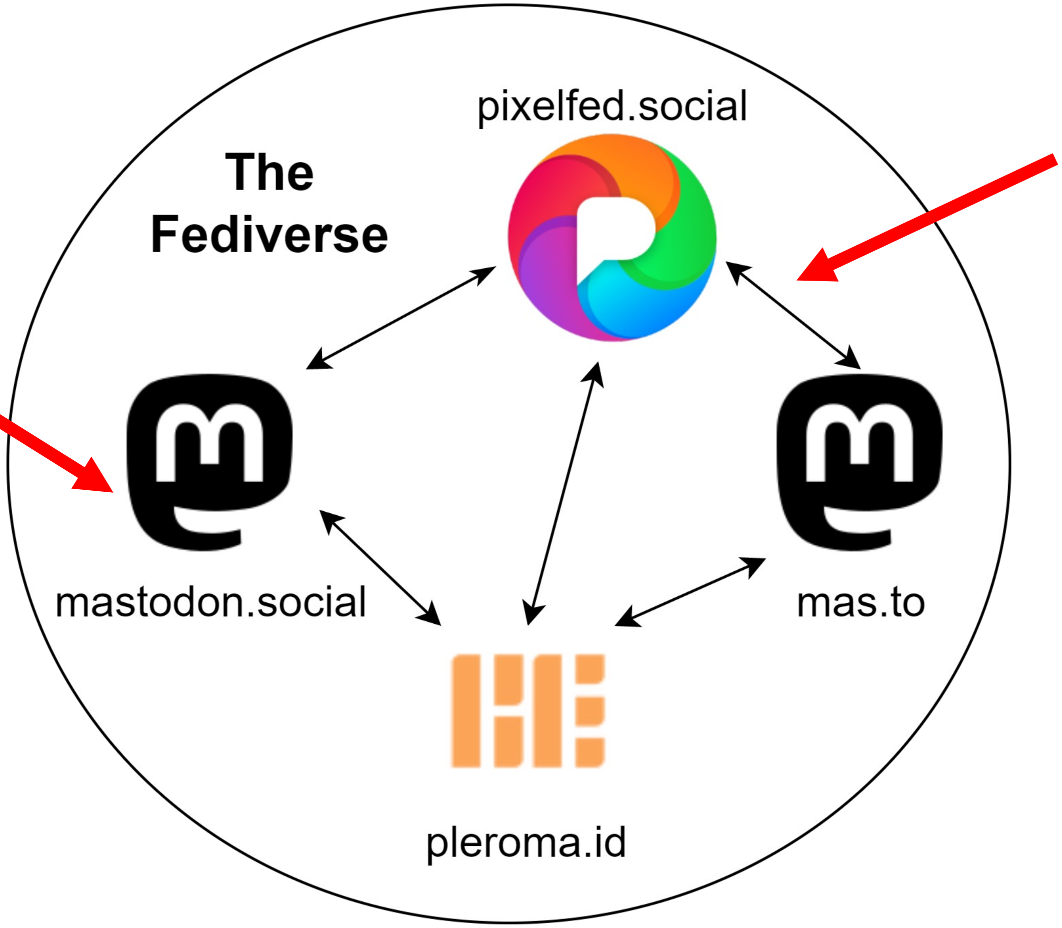
- Defined as "internet-based… persistent channel[s] of mass personal communication facilitating perceptions of interactions among users, deriving value primarily from user-generated content" (Carr & Hayes, 2014)
- Smith’s Honeycomb Model (Smith 2007)



Honeycomb Model
(Smith, 2007)

Open-source Software SMPs

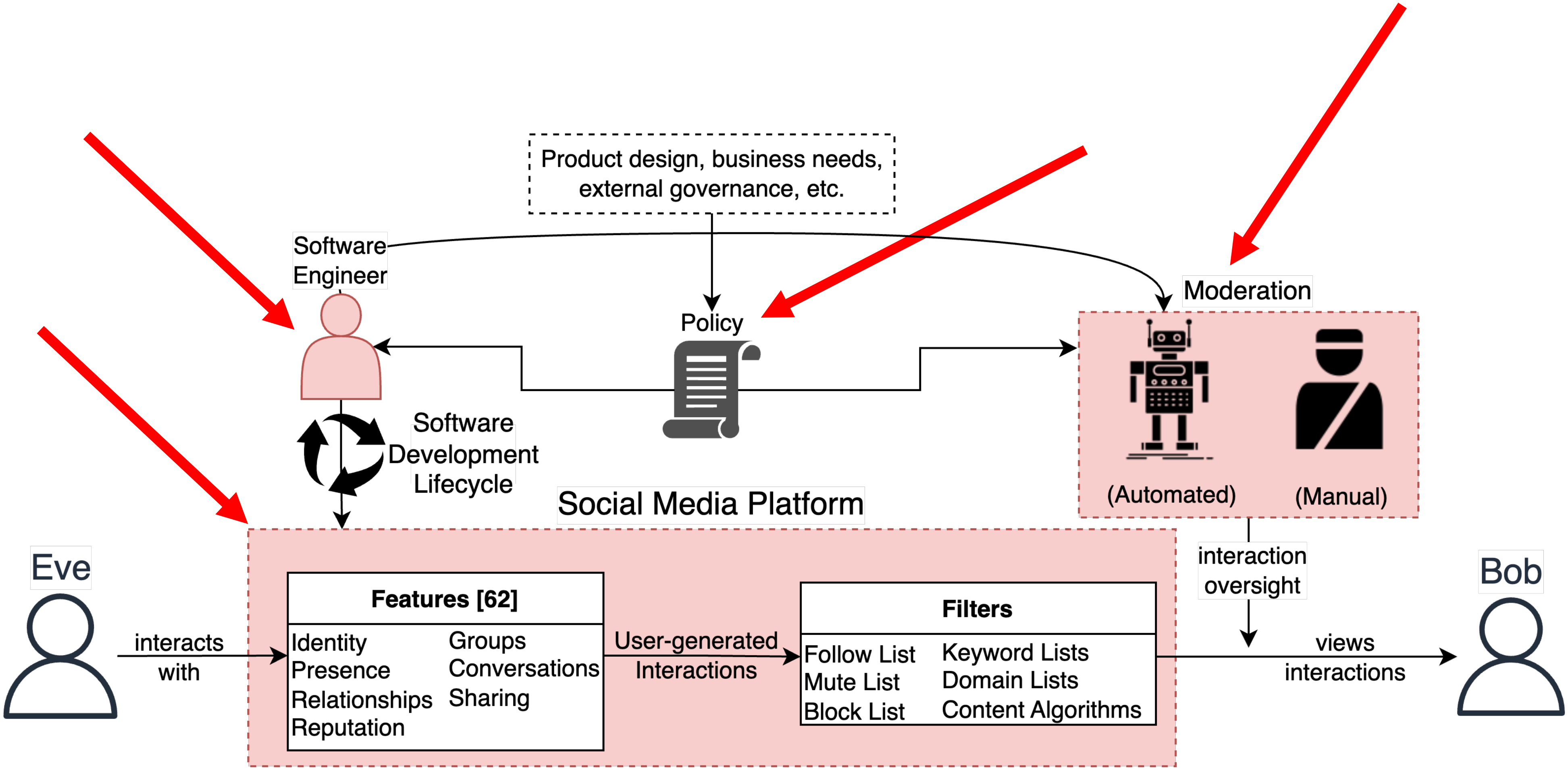
- Open-source Software (OSS) SMPs first appeared around 2010 and leveraged public sharing protocols
- The most popular OSS SMP, Mastodon, appeared in 2016 and mimics Twitter



Instance

ActivityPub Protocol

Social Media Platform (SMP): Context Diagram



Trust & Safety (T&S) and its Engineering

- 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)

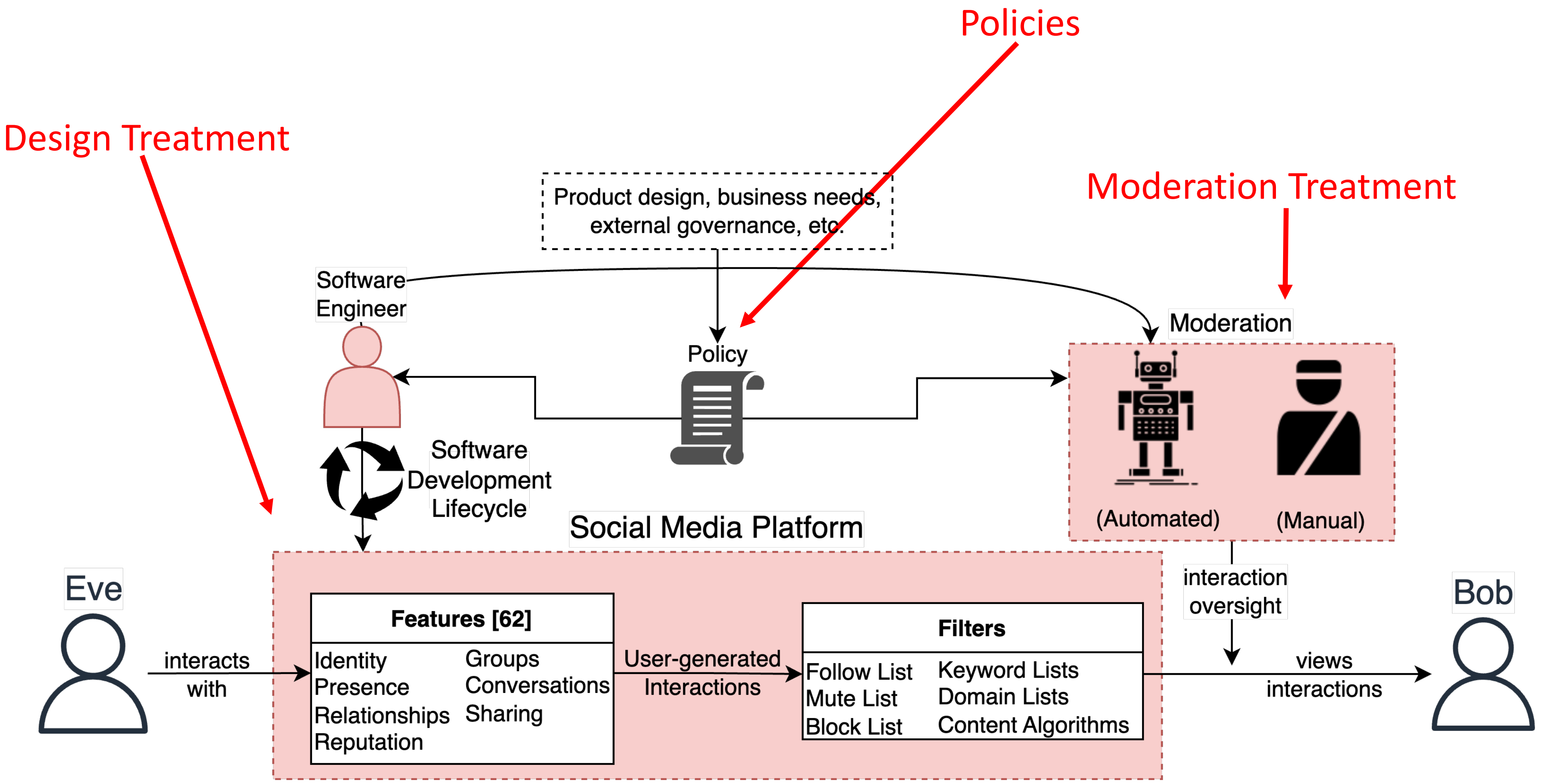
T&S in SMPs: A Risk Management View

- 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



Policies

Design Treatment

Moderation Treatment

Product design, business needs, external governance, etc.

Software Engineer

Policy

Moderation

Software Development Lifecycle

(Automated)

(Manual)

Social Media Platform

Eve

interacts with

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

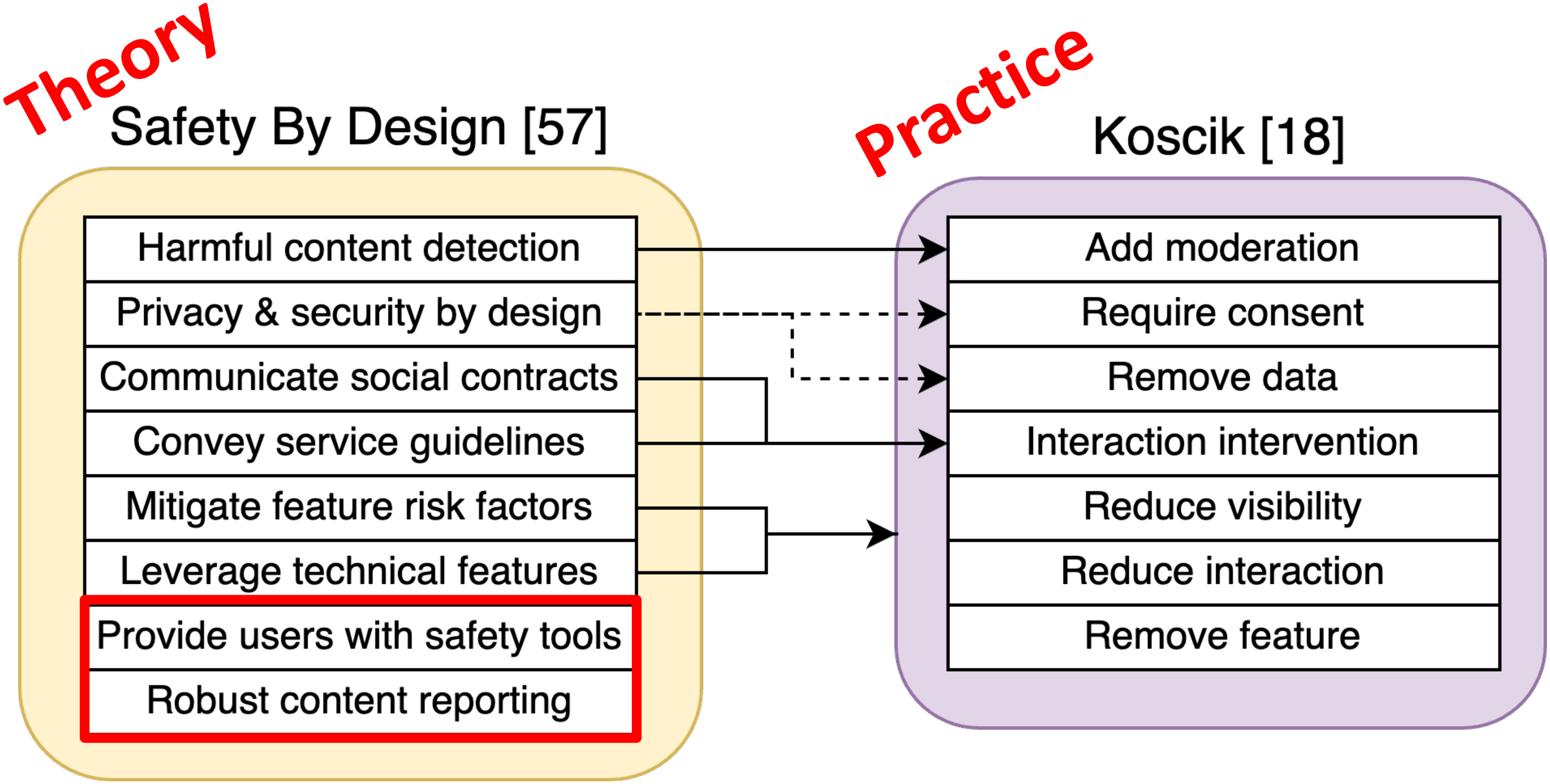
views interactions

Bob

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

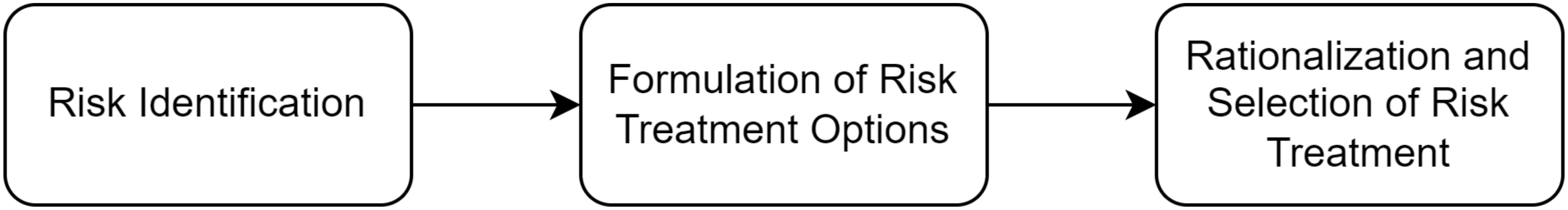
Risk Assessment

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 Identification
Formulation of Risk Treatment Options
Rationalization and Selection of Risk Treatment

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) #10384

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

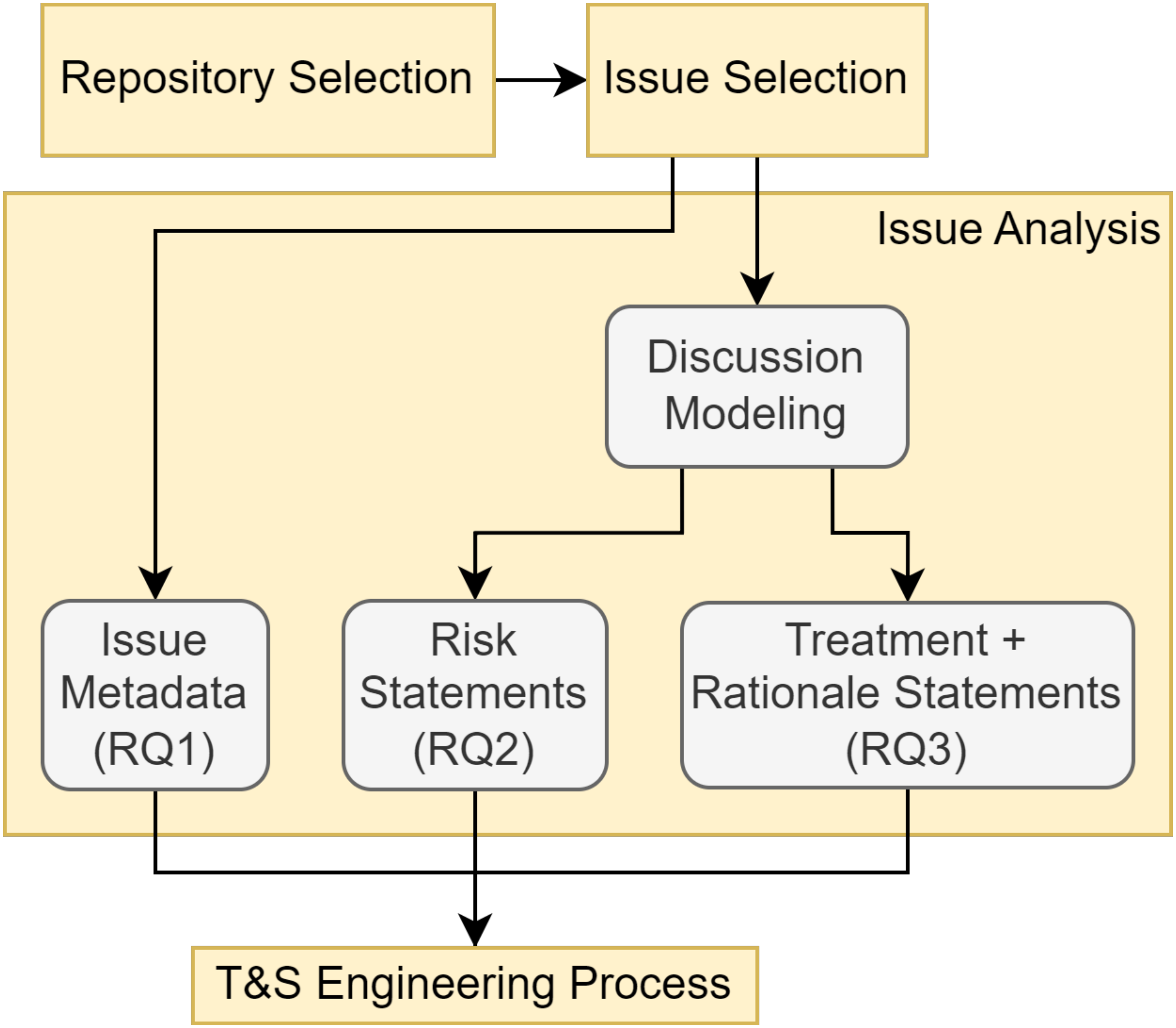
User B: many people are using limited bio space to do so but are not easy to access

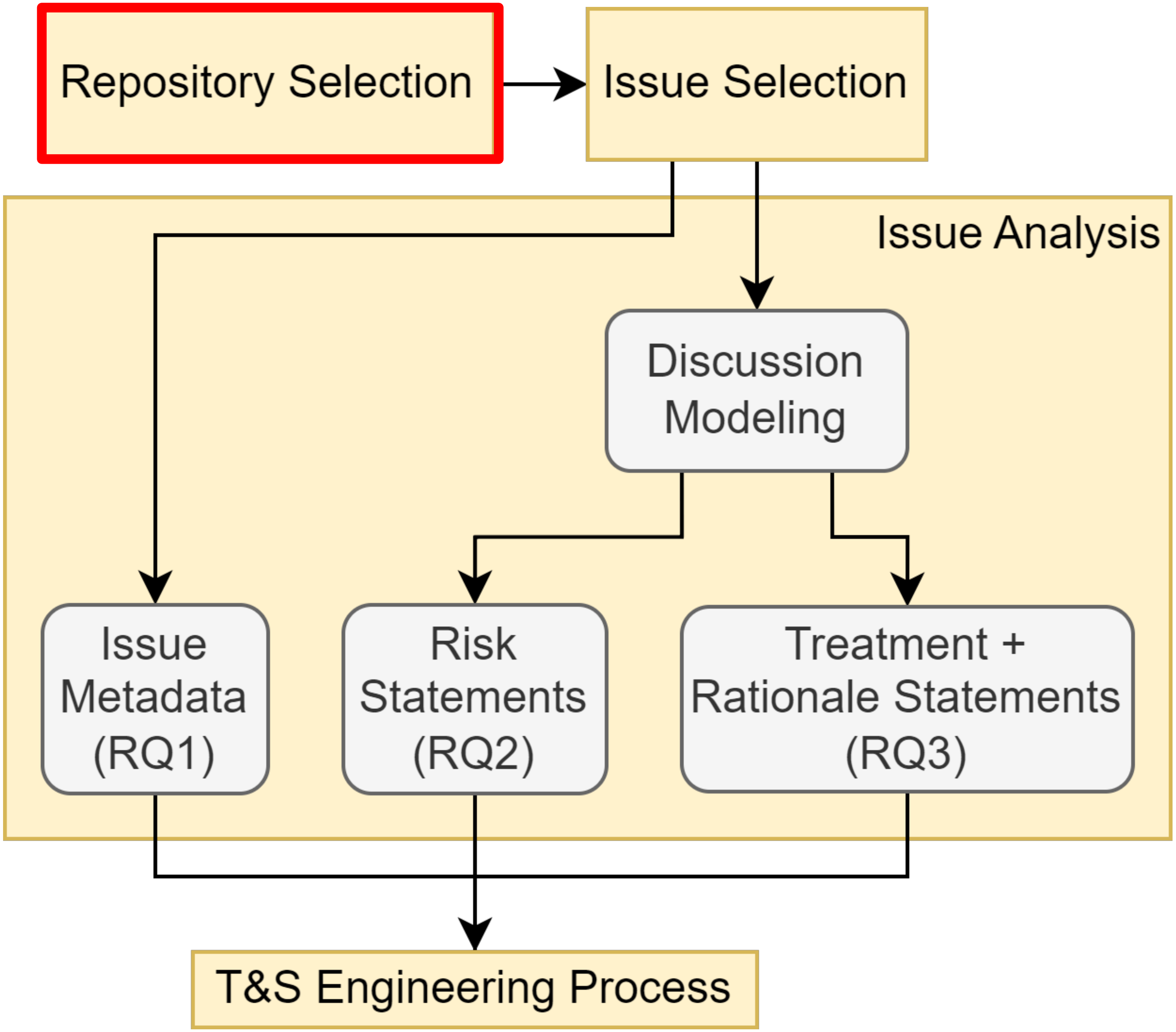
User C: If the message can be set by the user being replied to, it’s an avenue for abuse

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

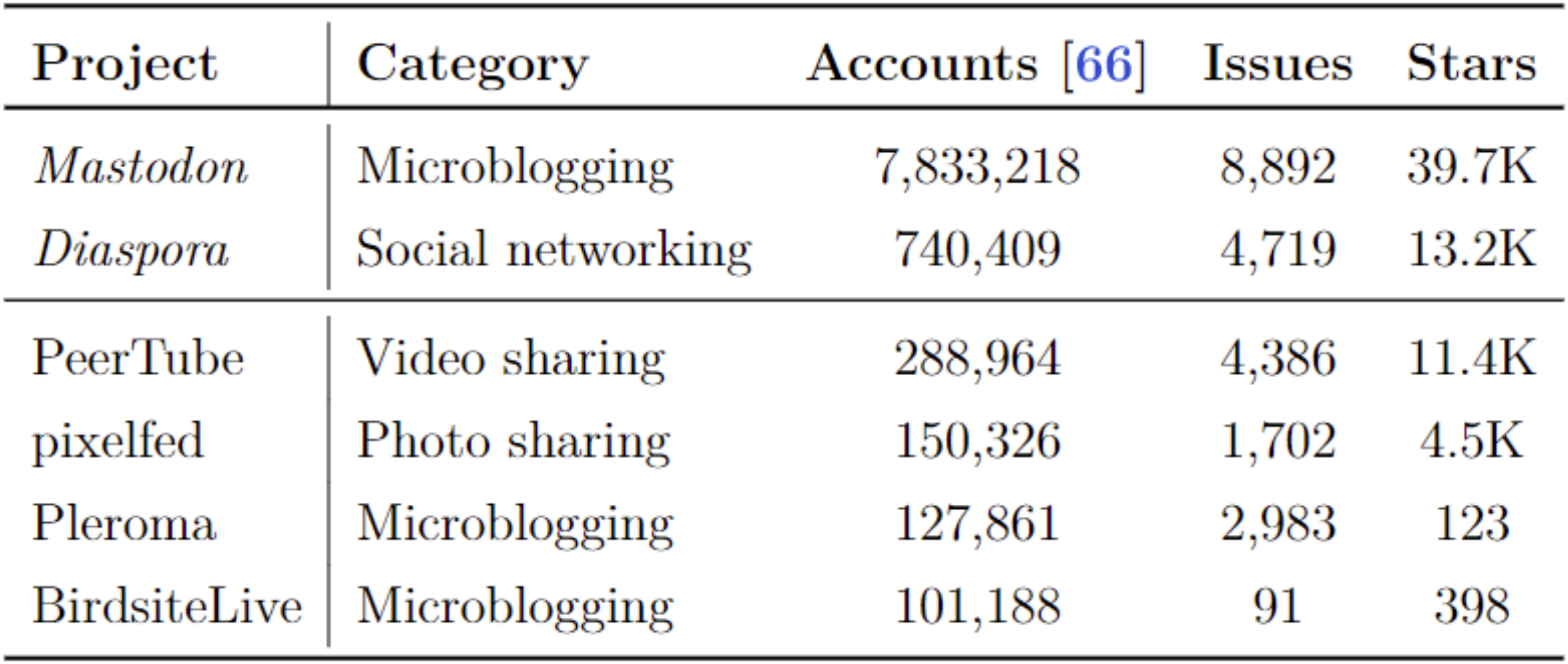
User C: annoying types could use a client that does not implement this feature

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


Repository Selection
Issue Selection
Issue Analysis
Discussion Modeling
Issue Metadata (RQ1)
Risk Statements (RQ2)
Treatment + Rationale Statements (RQ3)
T&S Engineering Process


Repository Selection
Issue Selection
Issue Analysis
Discussion Modeling
Issue Metadata (RQ1)
Risk Statements (RQ2)
Treatment + Rationale Statements (RQ3)
T&S Engineering Process

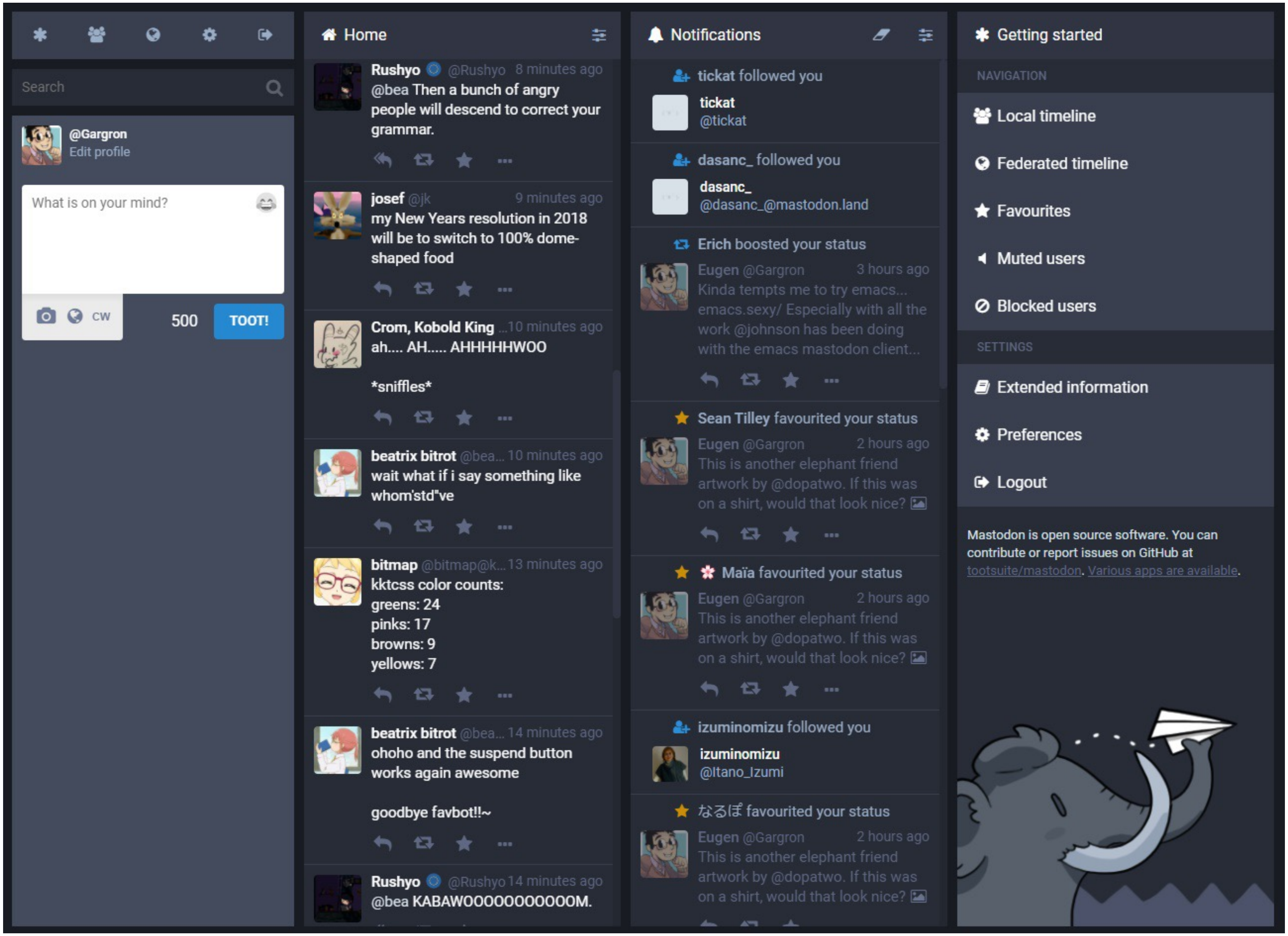
Repository Selection



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

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

Mastodon Homepage



Rushyo @Rushyo 8 minutes ago
@bea Then a bunch of angry people will descend to correct your grammar.

josef @jk 9 minutes ago
my New Years resolution in 2018 will be to switch to 100% dome-shaped food

Crom, Kobold King ...10 minutes ago
ah.... AH..... AHHHHHWOO

sniffles

beatrix bitrot @bea... 10 minutes ago
wait what if i say something like whom'std've

bitmap @bitmap@k... 13 minutes ago
kktcss color counts:
greens: 24
pinks: 17
browns: 9
yellows: 7

beatrix bitrot @bea... 14 minutes ago
ohoho and the suspend button works again awesome

goodbye favbot!!~

Rushyo @Rushyo 14 minutes ago
@bea KABAWOOOOOOOOOOOM.

tickat followed you
tickat
@tickat

dasanc_ followed you
dasanc_
@dasanc_@mastodon.land

Erich boosted your status
Eugen @Gargron 3 hours ago
Kinda tempts me to try emacs... emacs.sexy/ Especially with all the work @johnson has been doing with the emacs mastodon client...

Sean Tilley favourited your status
Eugen @Gargron 2 hours ago
This is another elephant friend artwork by @dopatwo. If this was on a shirt, would that look nice?

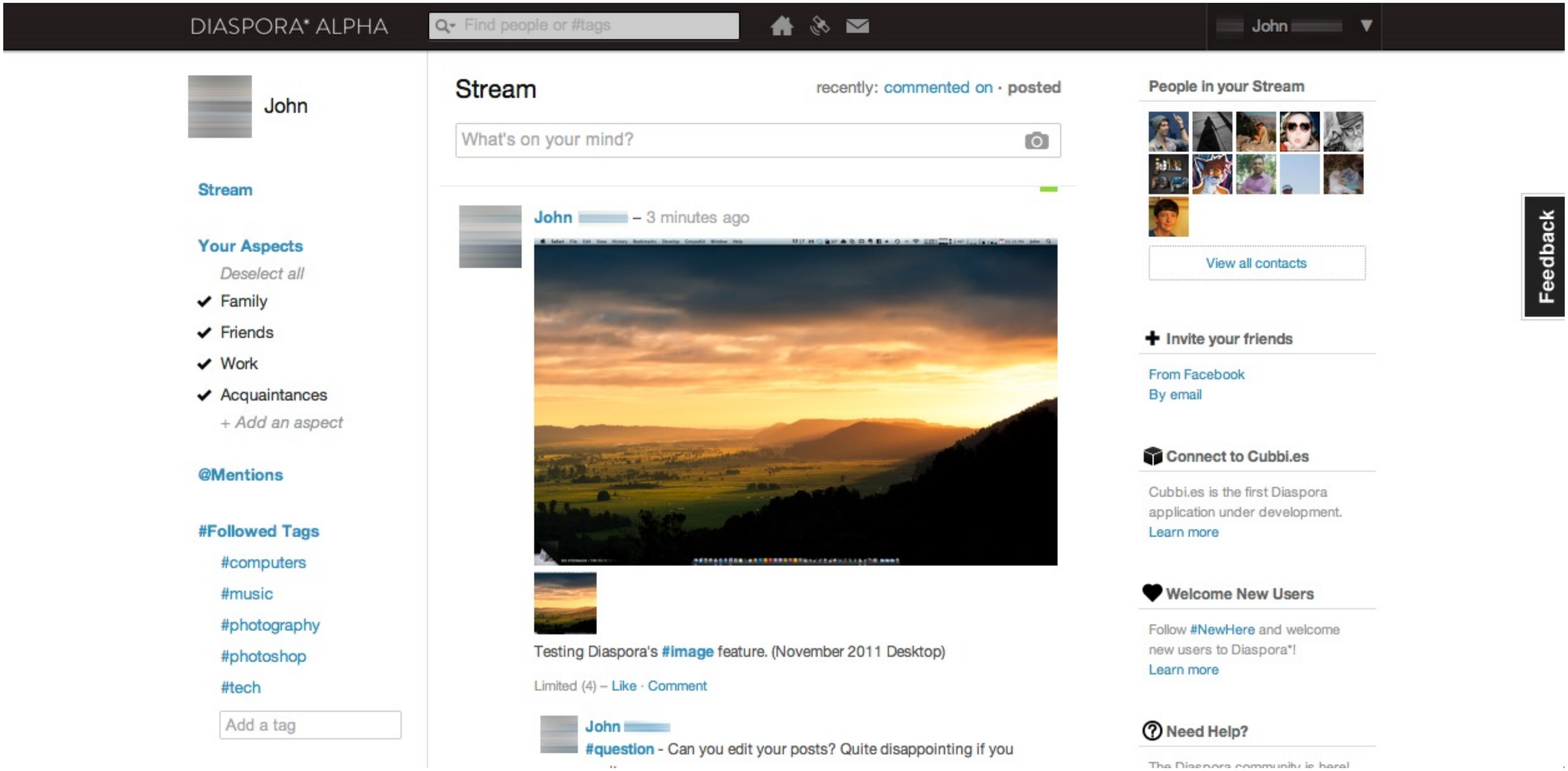
Maïa favourited your status
Eugen @Gargron 2 hours ago
This is another elephant friend artwork by @dopatwo. If this was on a shirt, would that look nice?

izuminomizu followed you
izuminomizu
@Itano_Izumi

なるほ favourited your status
Eugen @Gargron 2 hours ago
This is another elephant friend artwork by @dopatwo. If this was on a shirt, would that look nice?

Mastodon is open source software. You can contribute or report issues on GitHub at tootsuite/mastodon. Various apps are available.

Diaspora Homepage



DIASPORA* ALPHA

Find people or #tags

John [redacted] ▼

John

Stream

Your Aspects

Deselect all

- ✓ Family
- ✓ Friends
- ✓ Work
- ✓ Acquaintances

+ Add an aspect

@Mentions

#Followed Tags

- #computers
- #music
- #photography
- #photoshop
- #tech

Add a tag

Stream — recently: commented on · posted

What's on your mind?

John [redacted] – 3 minutes ago

Testing Diaspora's #image feature. (November 2011 Desktop)

Limited (4) – Like · Comment

John [redacted]

#question - Can you edit your posts? Quite disappointing if you

People in your Stream

View all contacts

✚ Invite your friends

From Facebook

By email

Connect to Cubbi.es

Cubbi.es is the first Diaspora application under development.

Learn more

Welcome New Users

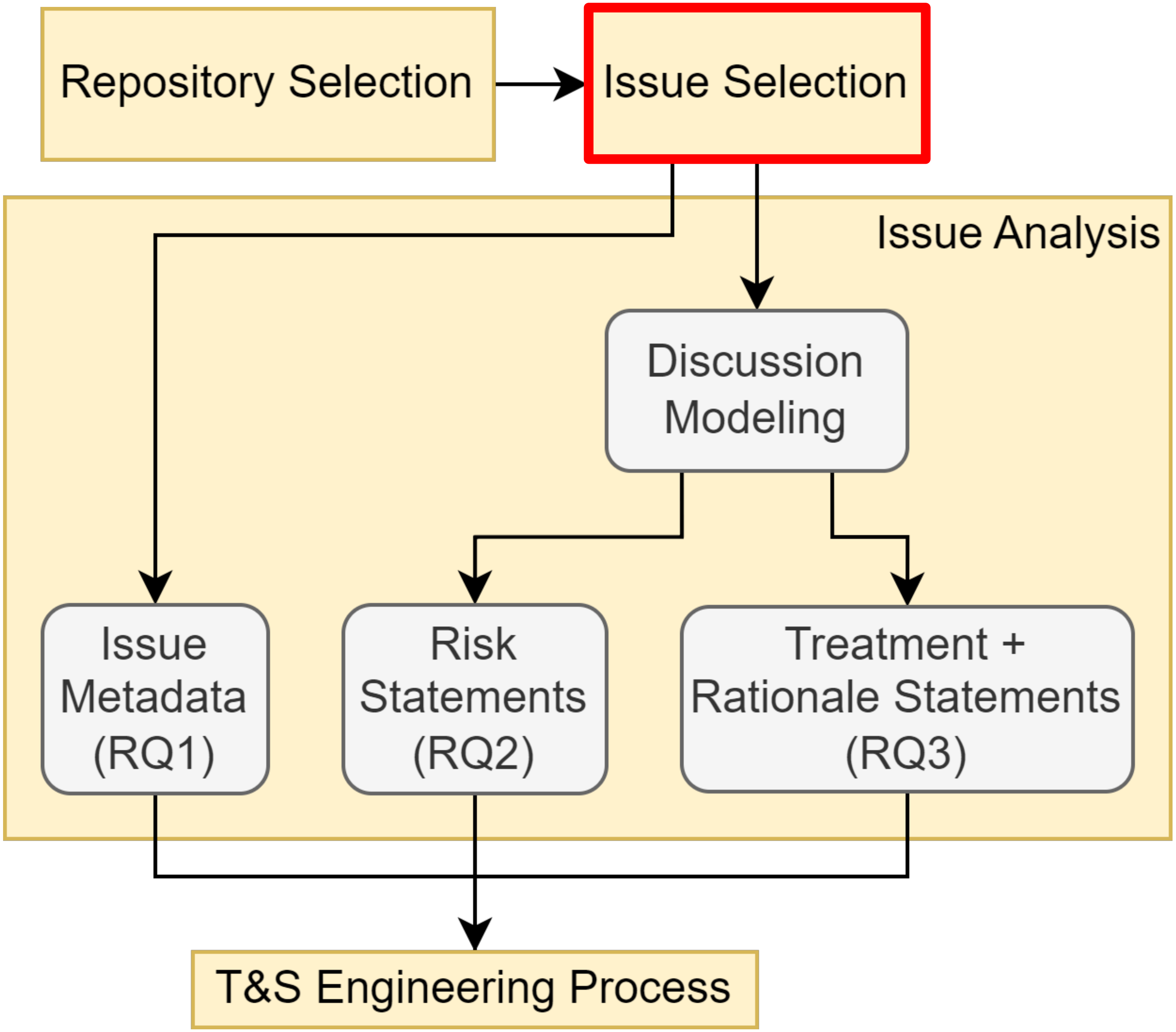
Follow #NewHere and welcome new users to Diaspora*!

Learn more

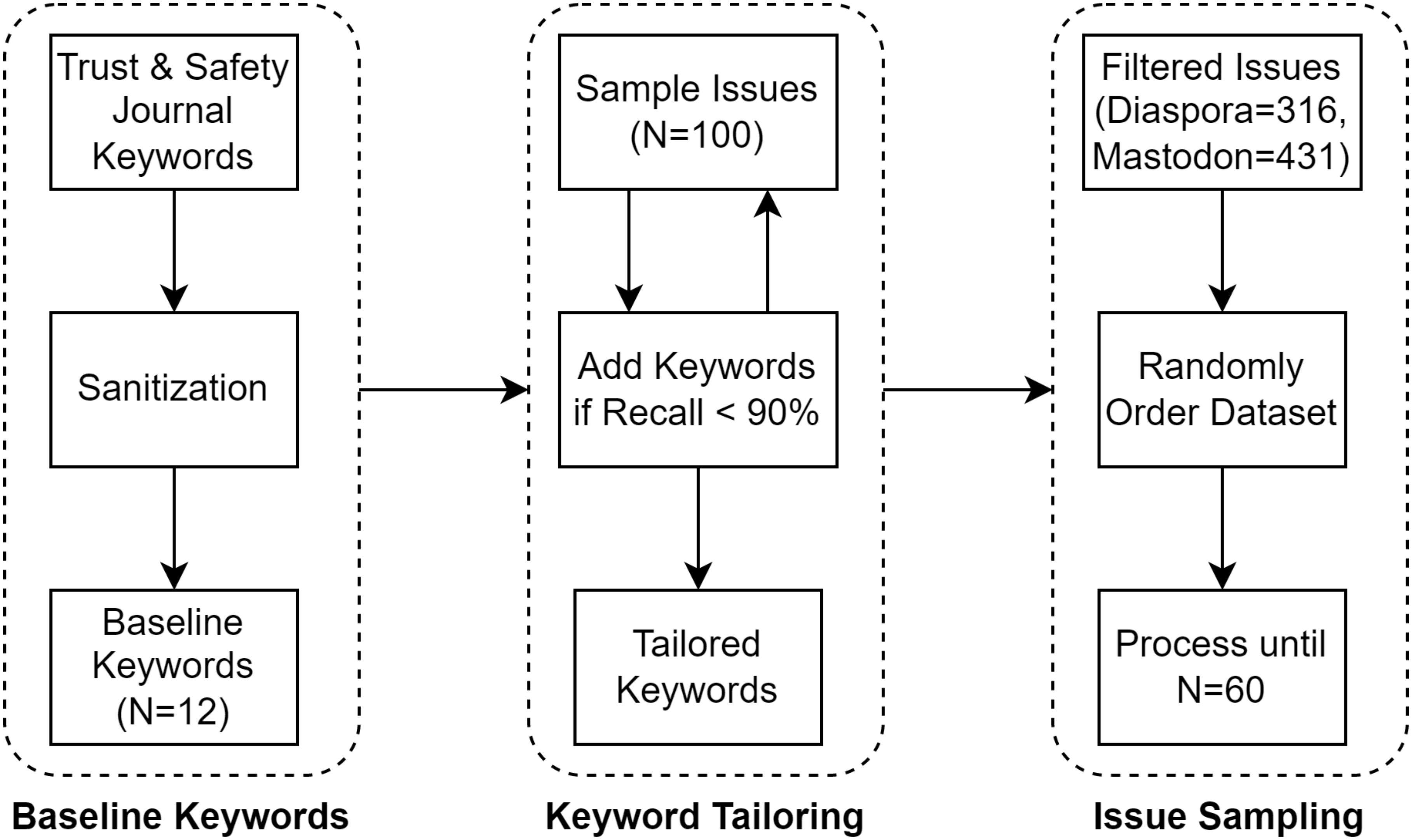
Need Help?

The Diaspora community is here!

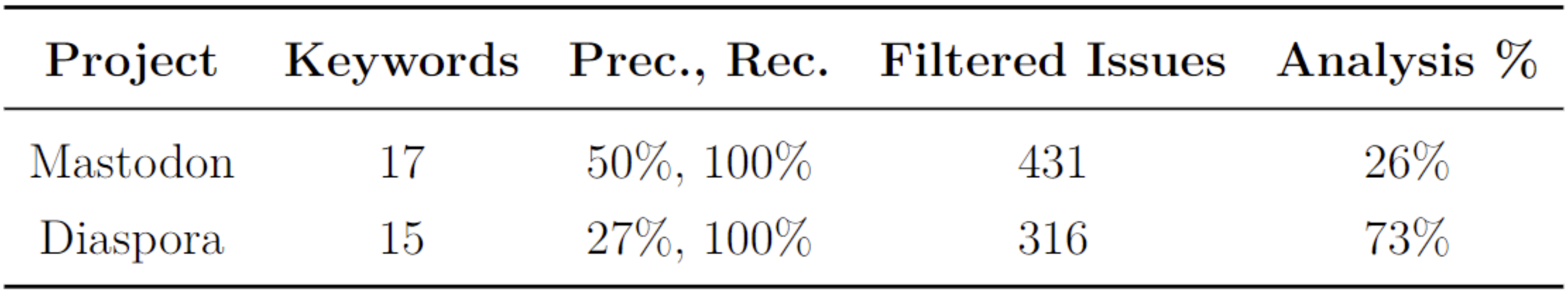
Feedback


Repository Selection
Issue Selection
Issue Analysis
Discussion Modeling
Issue Metadata (RQ1)
Risk Statements (RQ2)
Treatment + Rationale Statements (RQ3)
T&S Engineering Process

Issue Selection

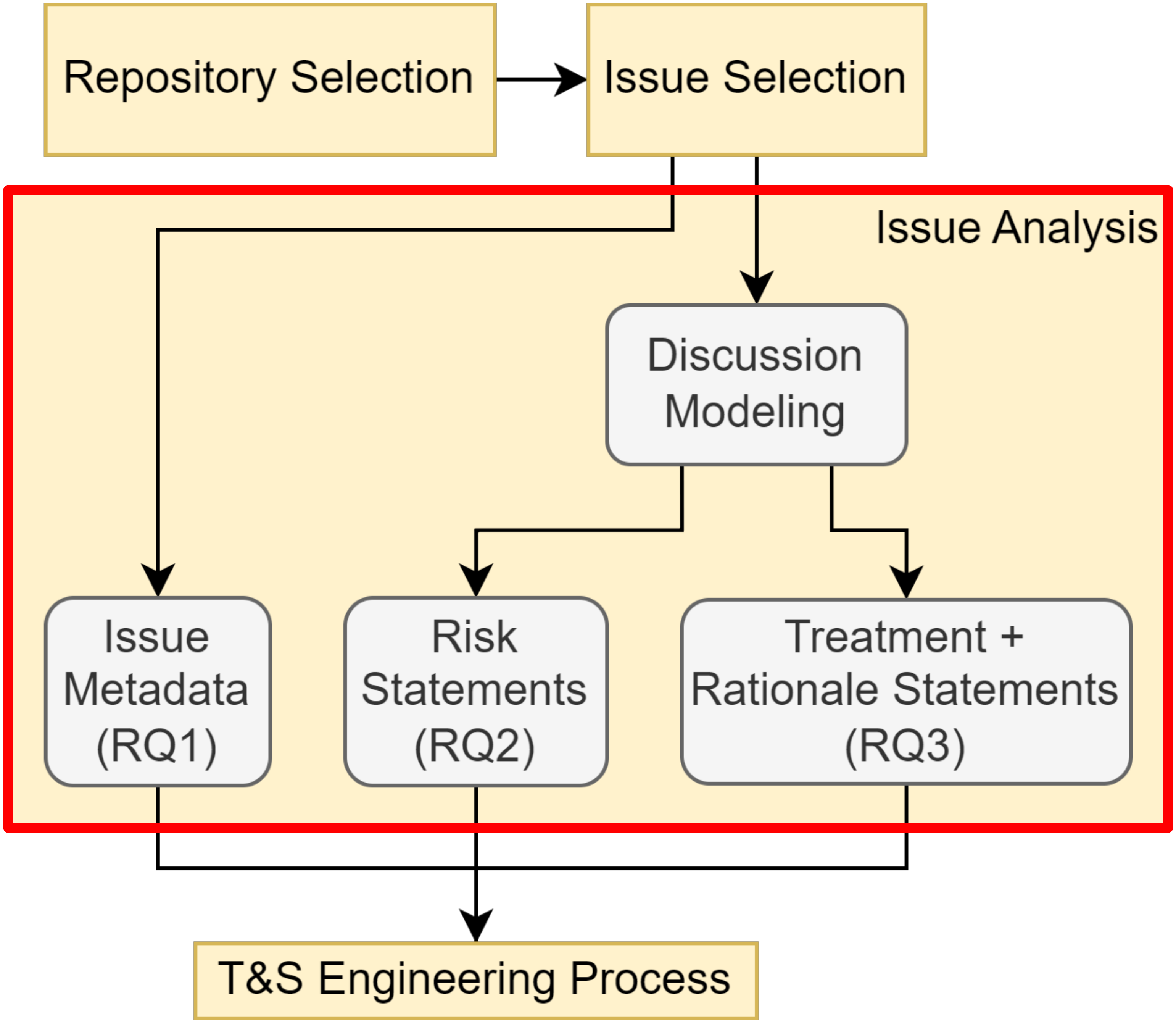


Issue Selection – Keyword Results

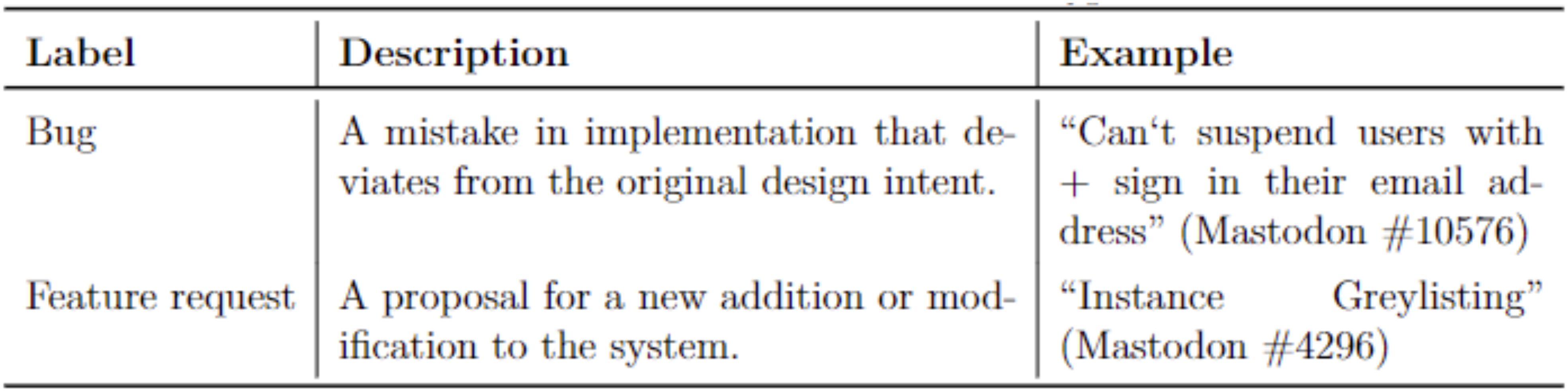


Project	Keywords	Prec., Rec.	Filtered Issues	Analysis %
Mastodon	17	50%, 100%	431	26%
Diaspora	15	27%, 100%	316	73%

SMP filtering results

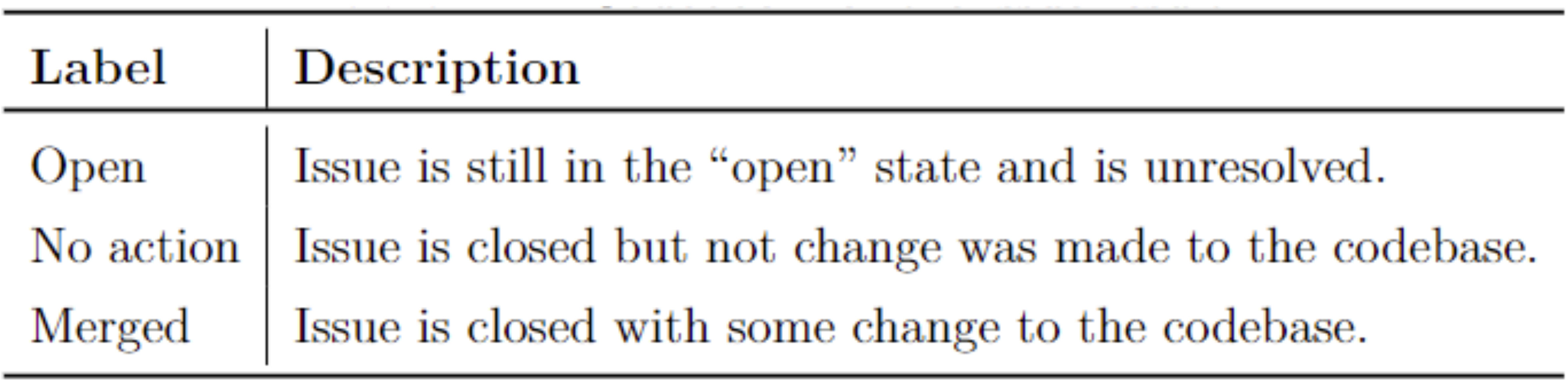

Repository Selection
Issue Selection
Issue Analysis
Discussion Modeling
Issue Metadata (RQ1)
Risk Statements (RQ2)
Treatment + Rationale Statements (RQ3)
T&S Engineering Process

Issue Analysis – Codebooks



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

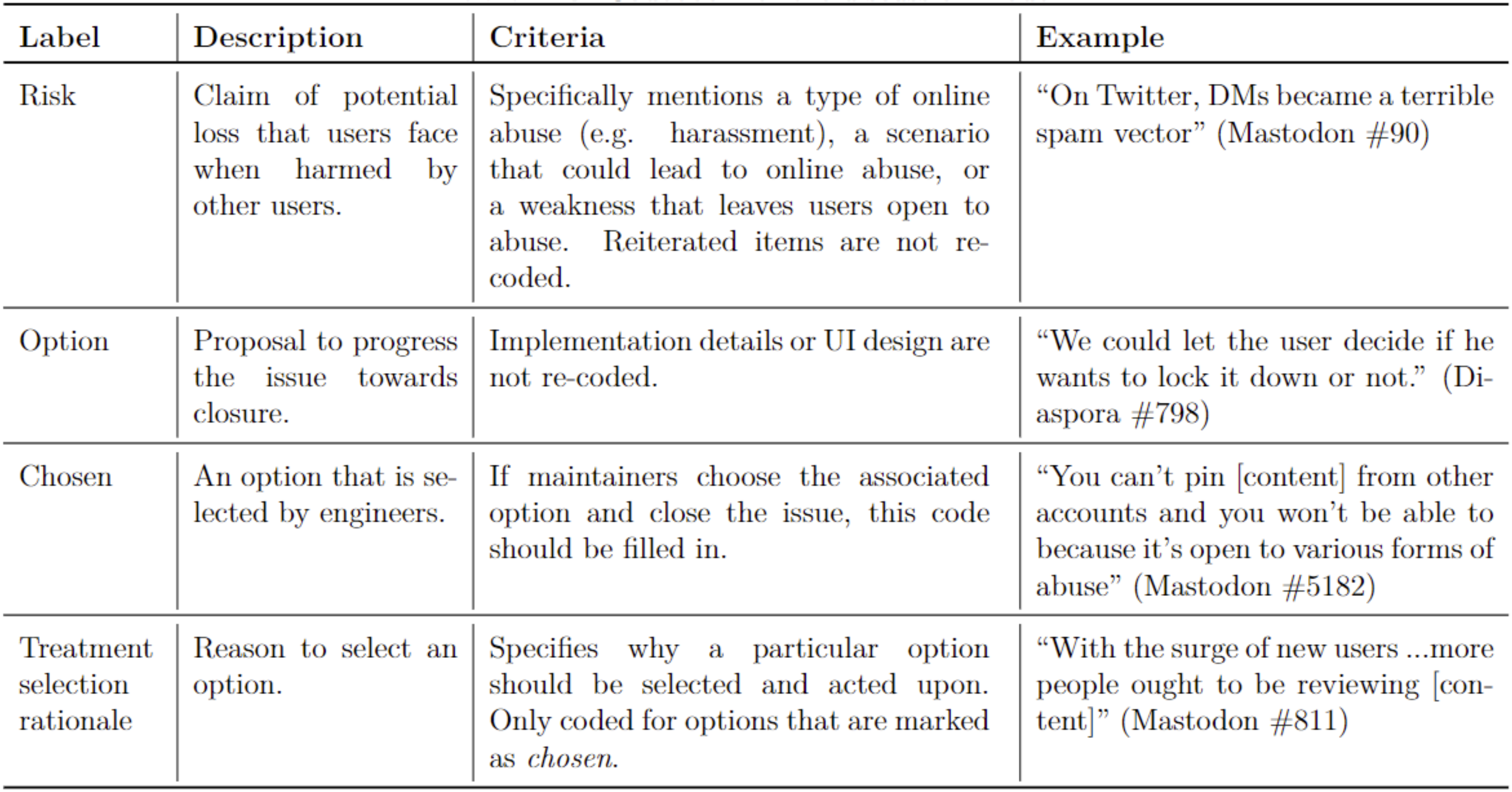
Codebook for *issue type*



Label	Description
Open	Issue is still in the "open" state and is unresolved.
No action	Issue is closed but not change was made to the codebase.
Merged	Issue is closed with some change to the codebase.

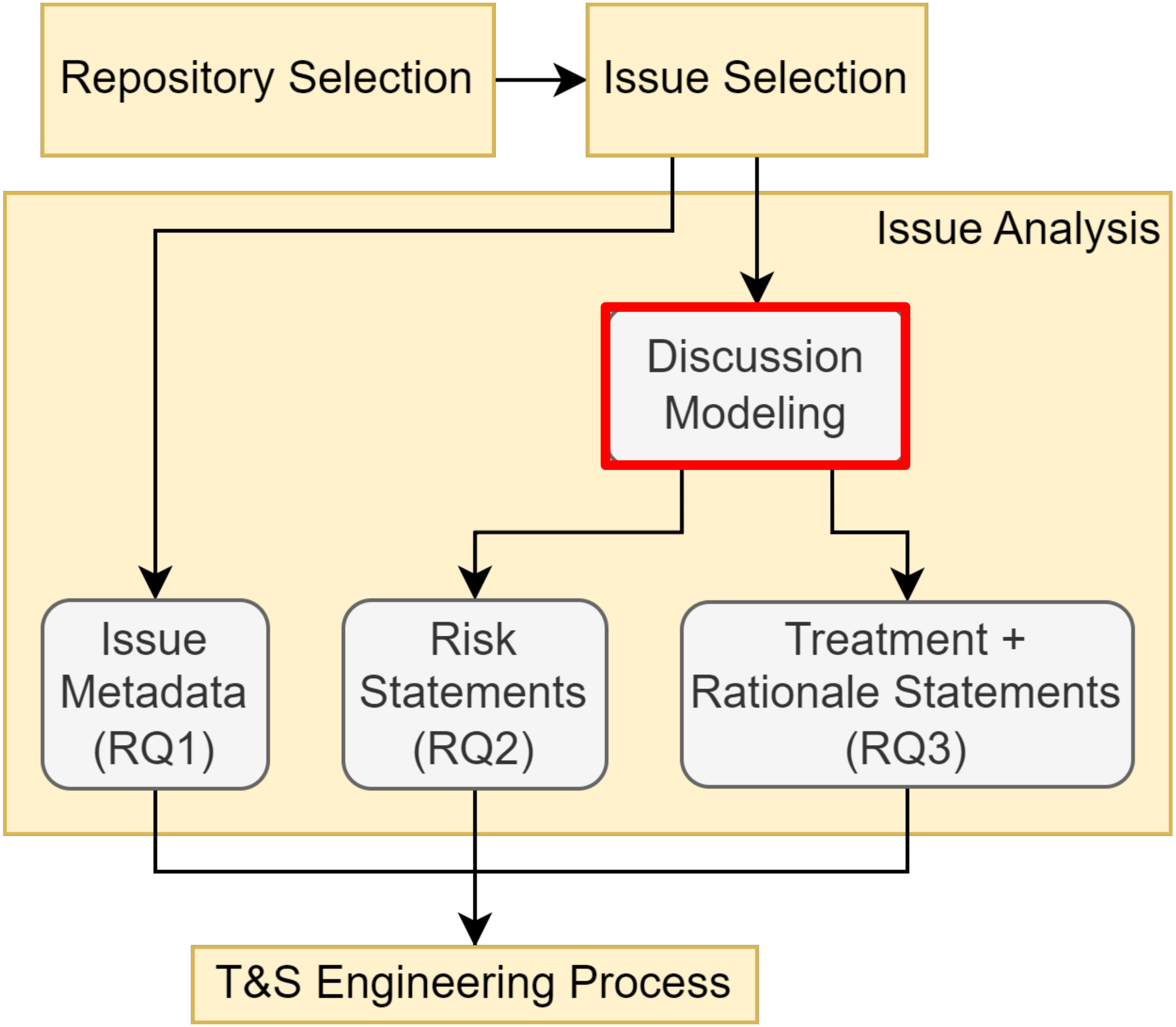
Codebook for *issue result*

Issue Analysis – Codebooks (cont.)



Label	Description	Criteria	Example
Risk	Claim of potential loss that users face when harmed by other users.	Specifically mentions a type of online abuse (e.g. harassment), a scenario that could lead to online abuse, or a weakness that leaves users open to abuse. Reiterated items are not re-coded.	"On Twitter, DMs became a terrible spam vector" (Mastodon #90)
Option	Proposal to progress the issue towards closure.	Implementation details or UI design are not re-coded.	"We could let the user decide if he wants to lock it down or not." (Diaspora #798)
Chosen	An option that is selected by engineers.	If maintainers choose the associated option and close the issue, this code should be filled in.	"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)
Treatment selection rationale	Reason to select an option.	Specifies why a particular option should be selected and acted upon. Only coded for options that are marked as *chosen*.	"With the surge of new users ...more people ought to be reviewing [content]" (Mastodon #811)

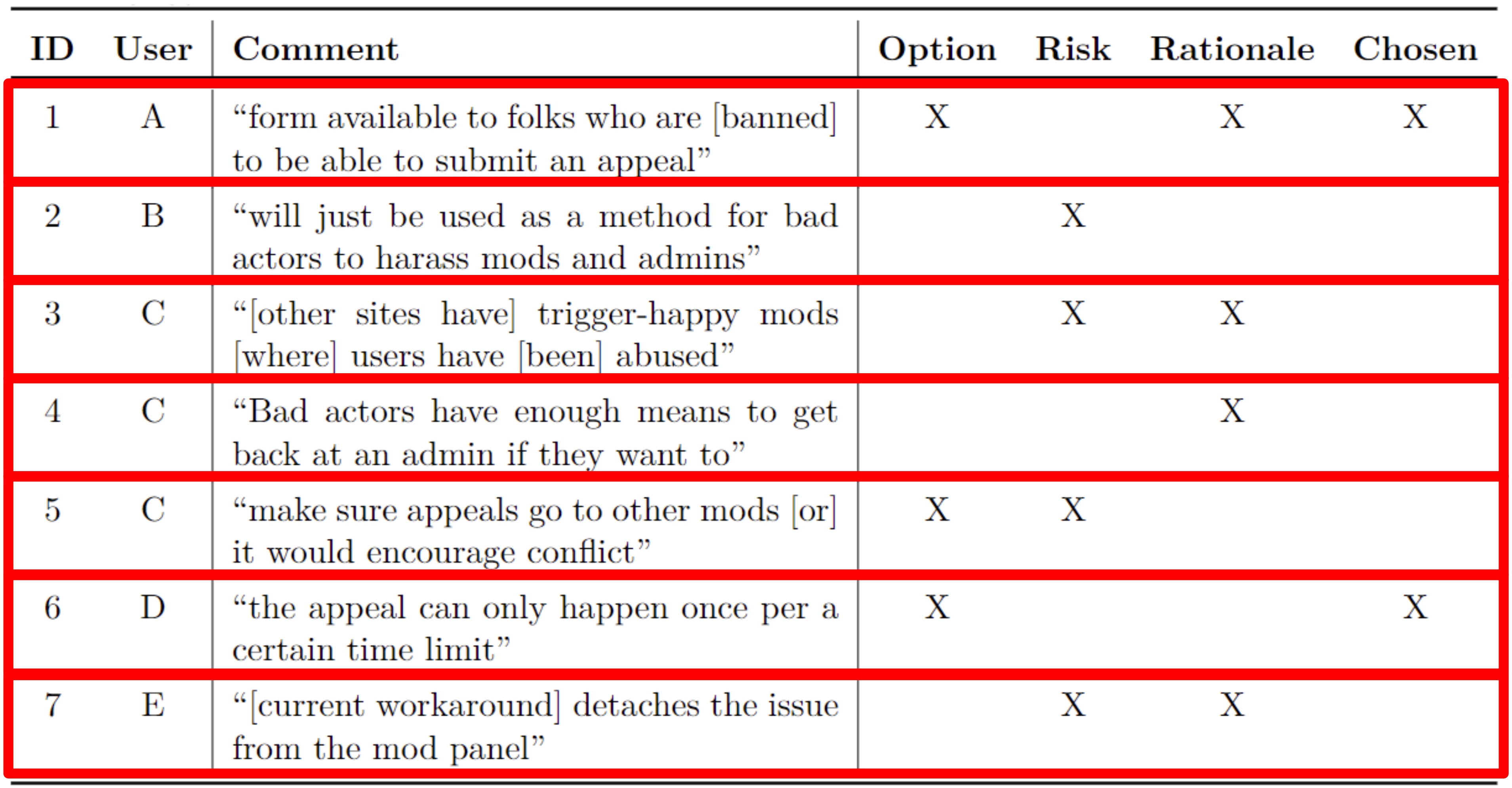
Codebook for discussion modeling


Repository Selection
Issue Selection
Issue Analysis
Discussion Modeling
Issue Metadata (RQ1)
Risk Statements (RQ2)
Treatment + Rationale Statements (RQ3)
T&S Engineering Process

Discussion Modeling Example 1

Appeals Function (Closed)

#9791

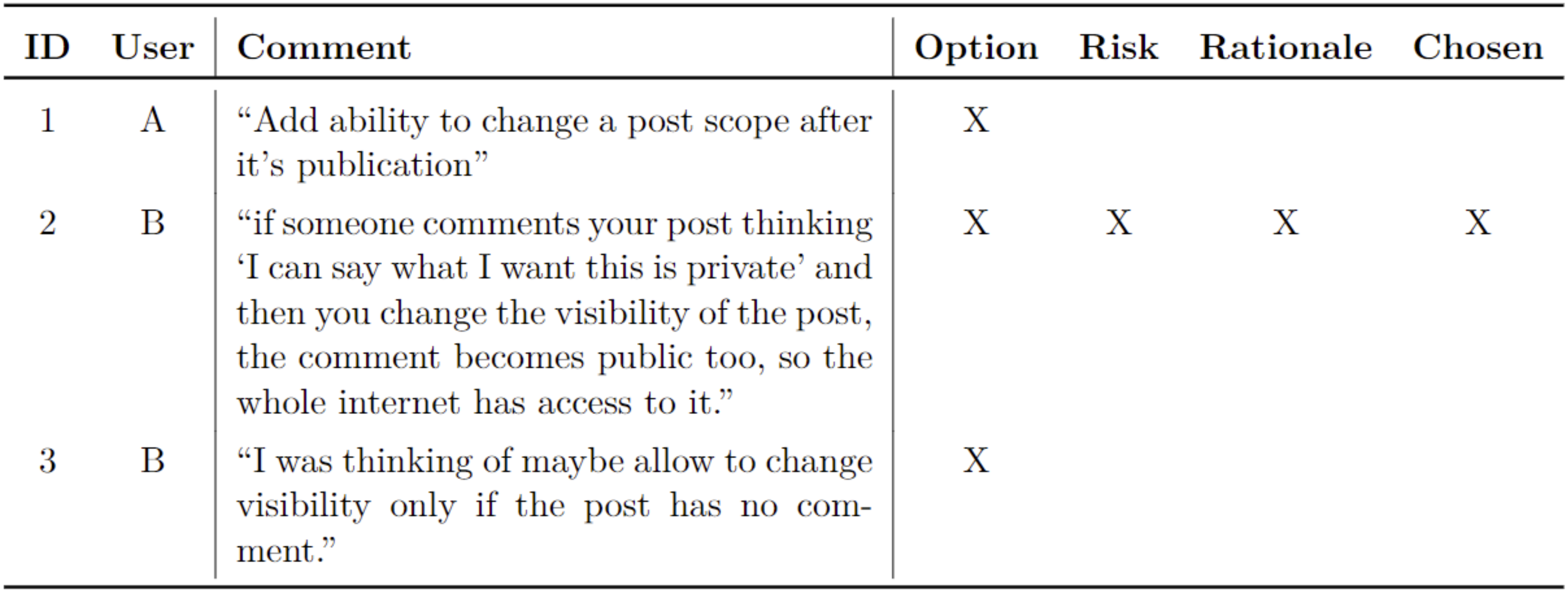


ID	User	Comment	Option	Risk	Rationale	Chosen
1	A	"form available to folks who are [banned] to be able to submit an appeal"	X		X	X
2	B	"will just be used as a method for bad actors to harass mods and admins"		X		
3	C	"[other sites have] trigger-happy mods [where] users have [been] abused"		X	X	
4	C	"Bad 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	

Discussion Modeling Example 2

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

#4664



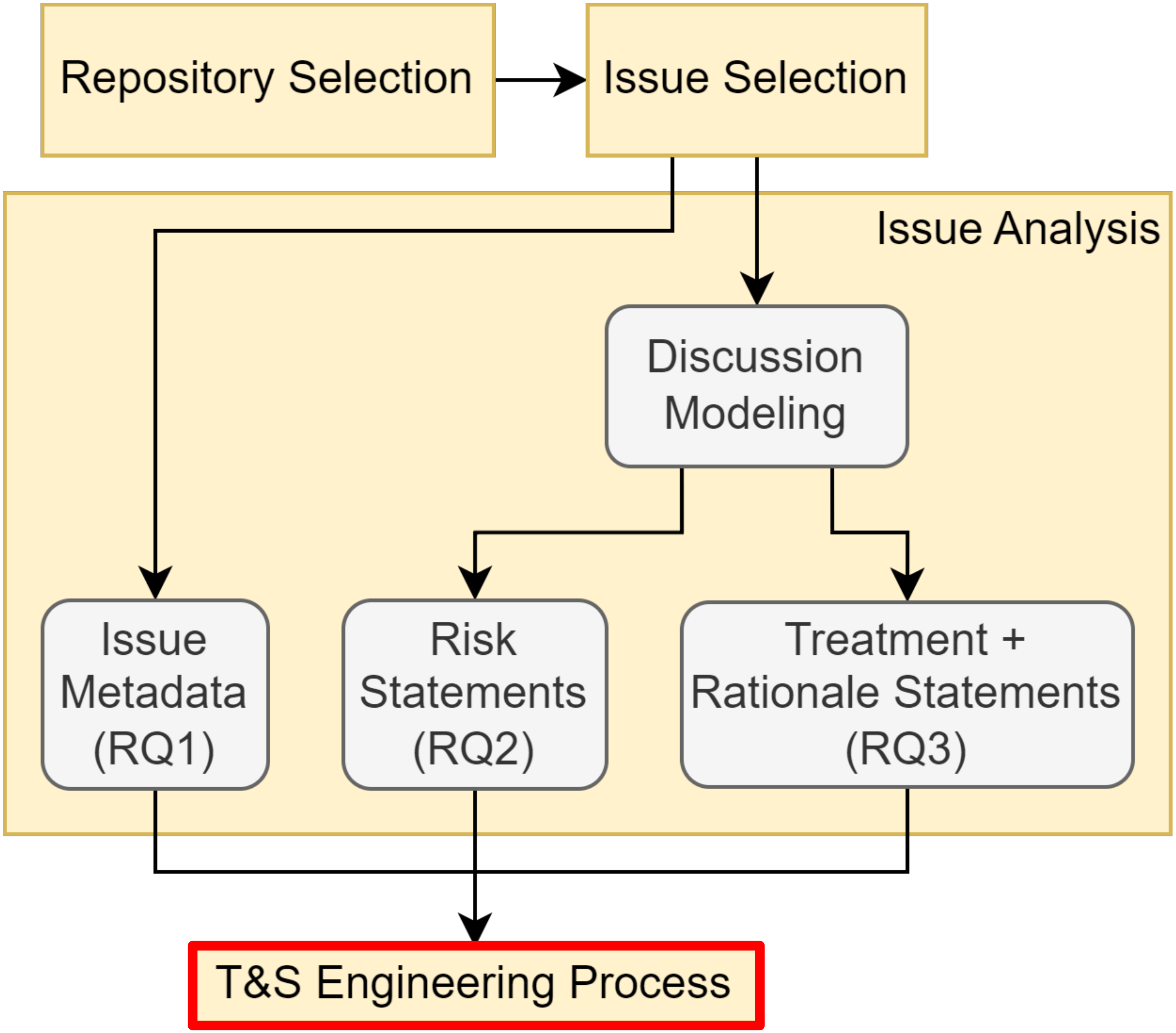
ID	User	Comment	Option	Risk	Rationale	Chosen
1	A	“Add ability to change a post scope after it’s publication”	X			
2	B	“if someone comments your post thinking ‘I can say what I want this is private’ and then you change the visibility of the post, the comment becomes public too, so the whole internet has 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			

Issue Analysis – Development of Taxonomies

Study Data	Related Work	Taxonomy
General issue topics	Honeycomb model (Smith, 2007)	SMP feature taxonomy
Risk sentences	Hate & harassment taxonomy (Thomas et al., 2021)	T&S risk taxonomy T&S threat actor taxonomy
Option sentences	Abuse vector treatment taxonomy (Koscik, 2018)	T&S Engineering pattern taxonomy
Rationale sentences	Software quality rationale taxonomy (Ko & Chilana, 2011)	T&S treatment selection rationale taxonomy

Issue Analysis – Inter-rater agreement

Study Data	Agreement method
Sentence codings	*Risk* κ = 0.89, *Option* κ = 1.0, *Rationale* κ = 1.0, *Chosen* κ = 1.0
SMP feature taxonomy	Did not pursue inter-rater agreement.
T&S Engineering pattern taxonomy	Random sample of 12/119. $\kappa = 0.73$
Rationale taxonomy	Random sample of 10%. $\kappa = 0.81$
Risk taxonomy	Collaborated on all categorizations.

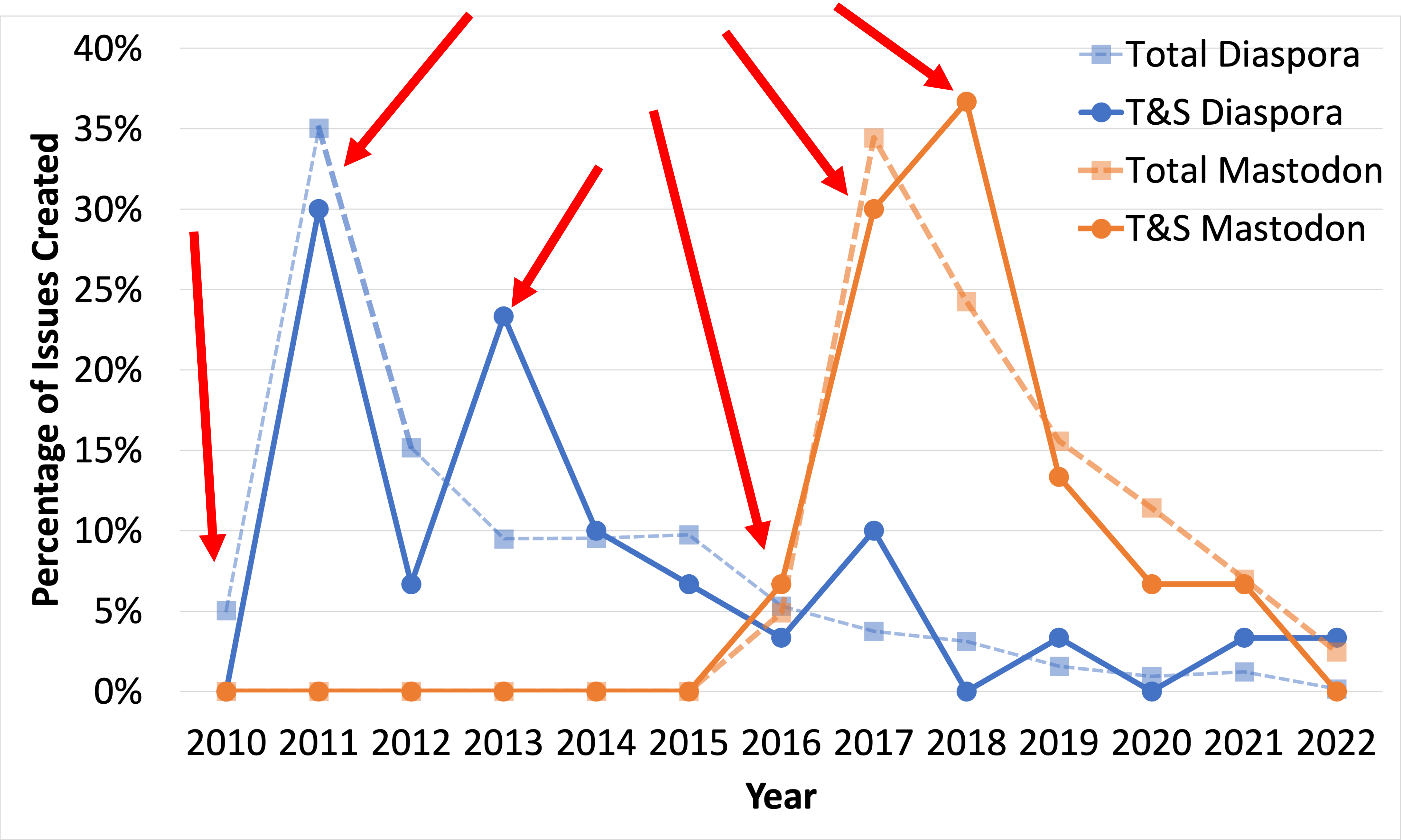

Repository Selection
Issue Selection
Issue Analysis
Discussion Modeling
Issue Metadata (RQ1)
Risk Statements (RQ2)
Treatment + Rationale Statements (RQ3)
T&S Engineering Process

Outline

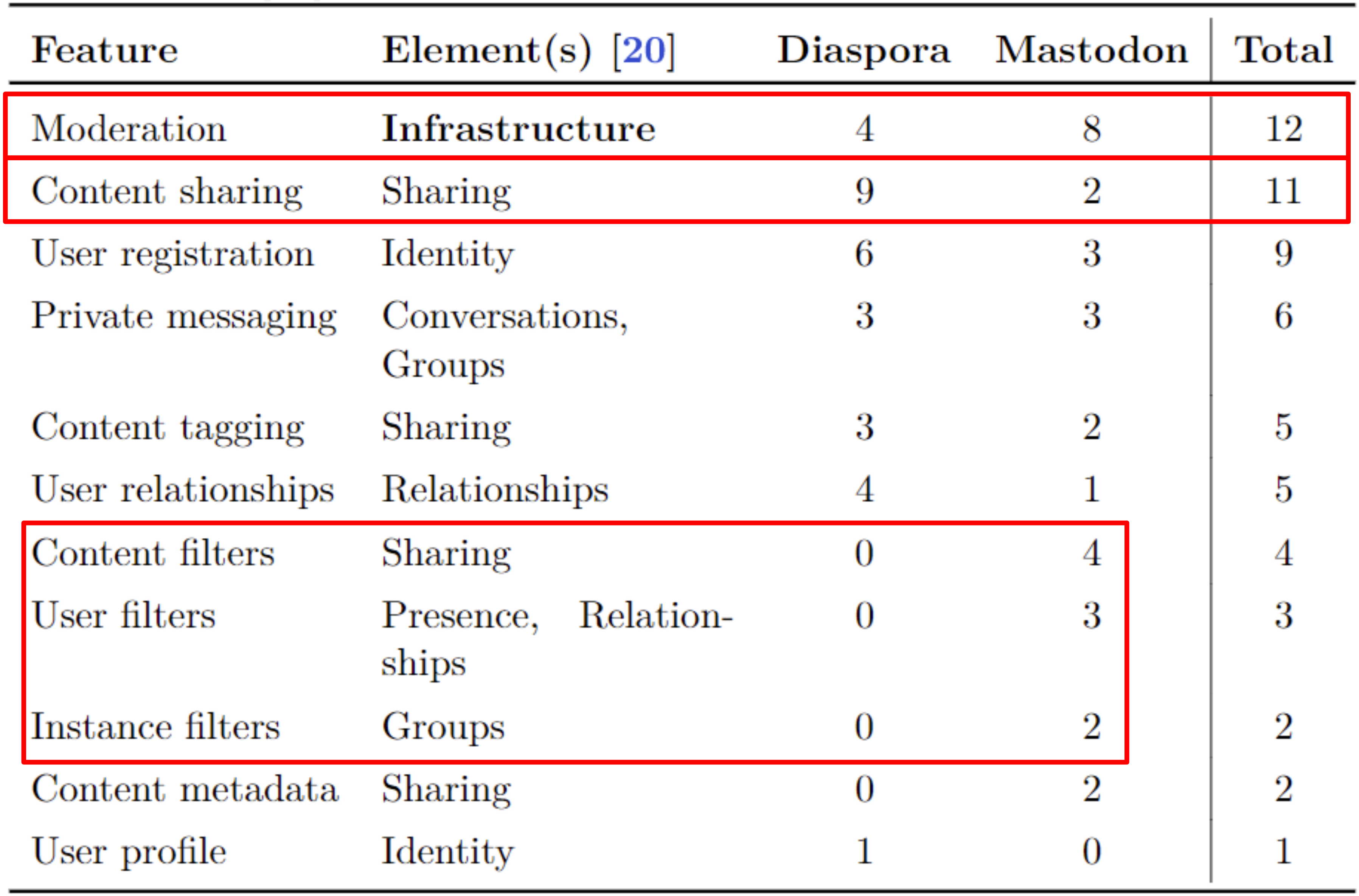
- Background
- Research Questions
- Methodology
- Results
- Discussion
- Threats to Validity

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

T&S issue curve is distinct



SMP Feature Breakdown



Feature	Element(s) [20]	Diaspora	Mastodon	Total
Moderation	**Infrastructure**	4	8	12
Content sharing	Sharing	9	2	11
User registration	Identity	6	3	9
Private messaging	Conversations, Groups	3	3	6
Content tagging	Sharing	3	2	5
User relationships	Relationships	4	1	5
Content filters	Sharing	0	4	4
User filters	Presence, Relationships	0	3	3
Instance filters	Groups	0	2	2
Content metadata	Sharing	0	2	2
User profile	Identity	1	0	1

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?

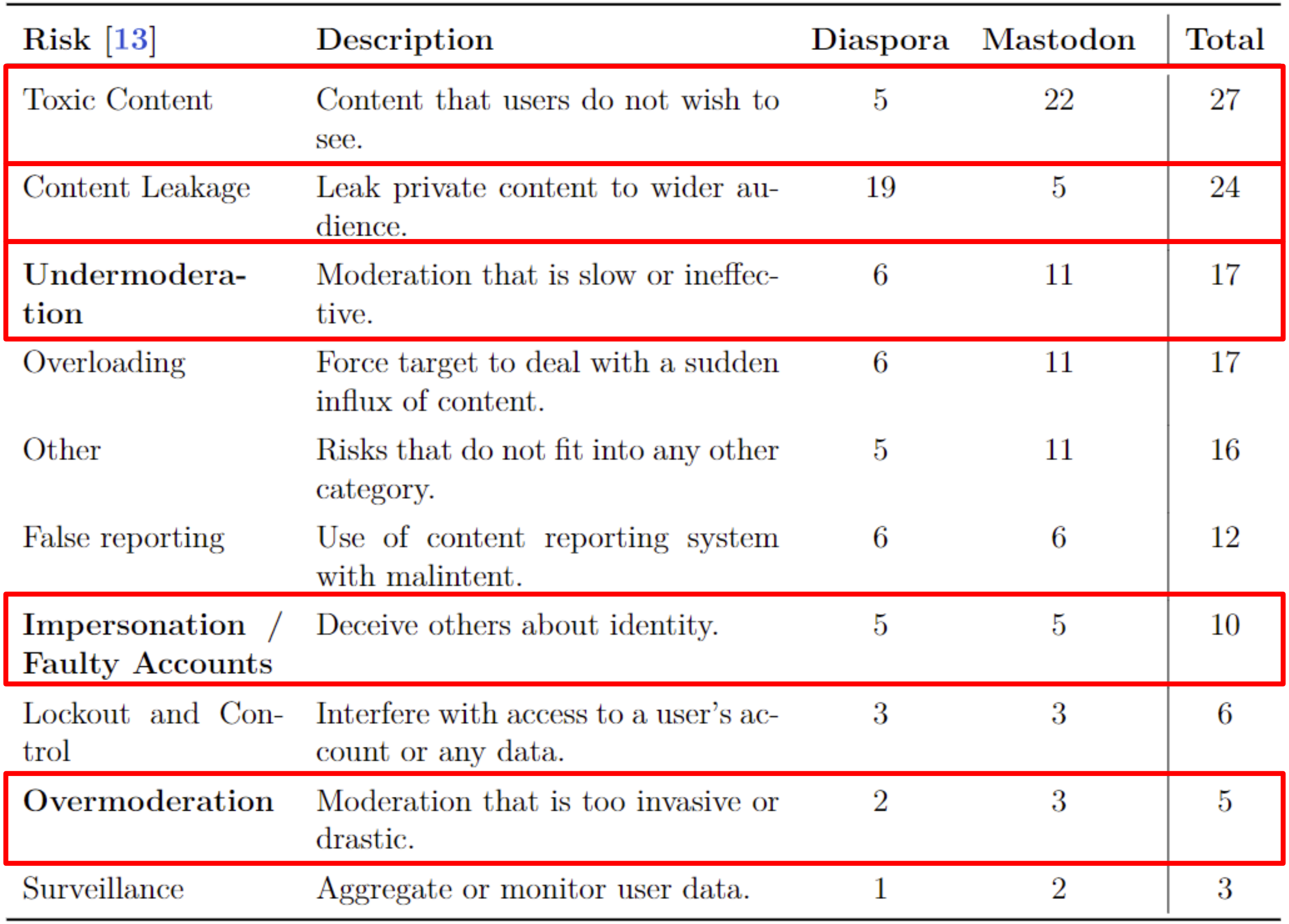
Threat Actor Analysis

- 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.

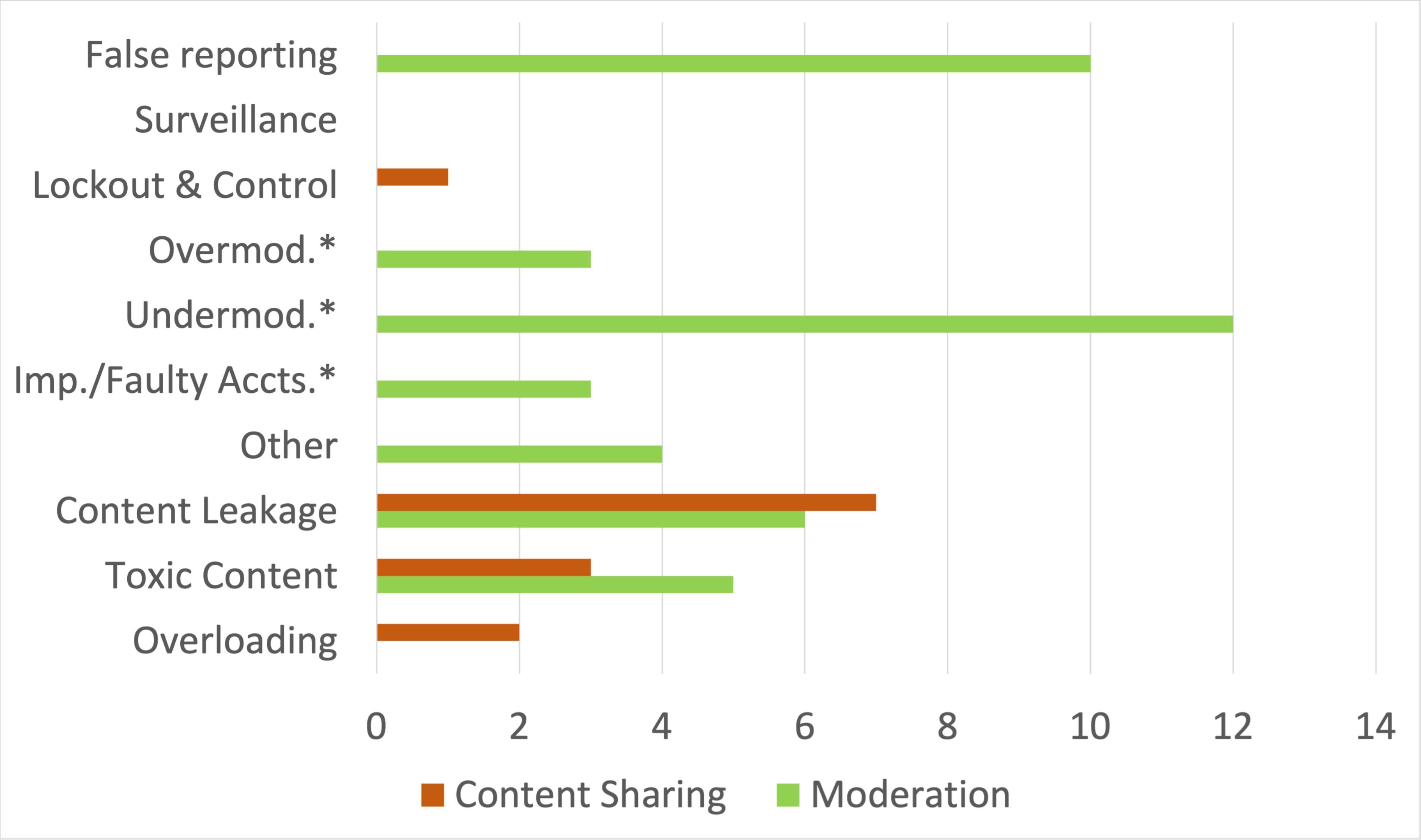
- 20% of risk statements identify moderator as a threat actor.

T&S Risk Taxonomy



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

Risk Landscapes – Moderation is hard

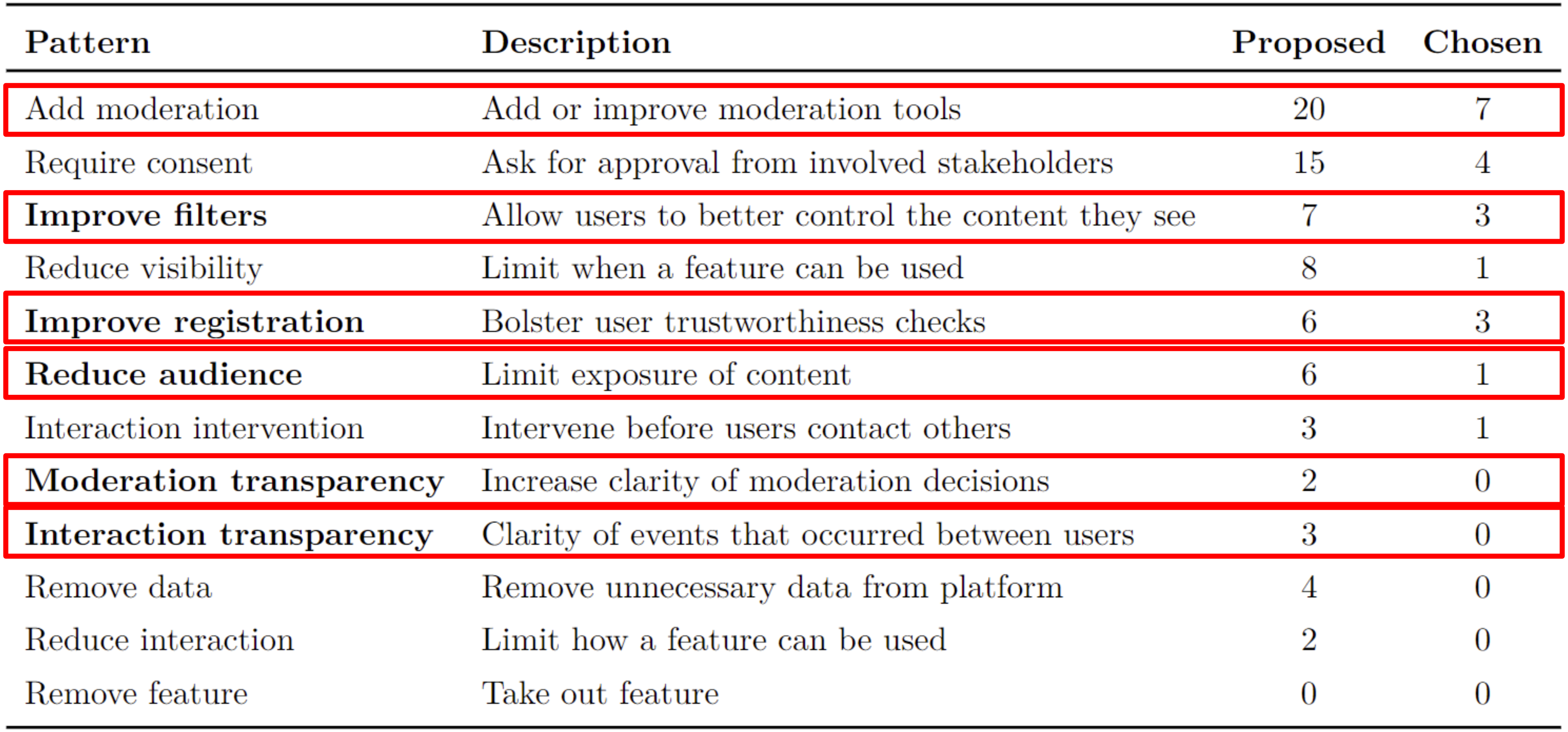


RQ2: Key Findings

- 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

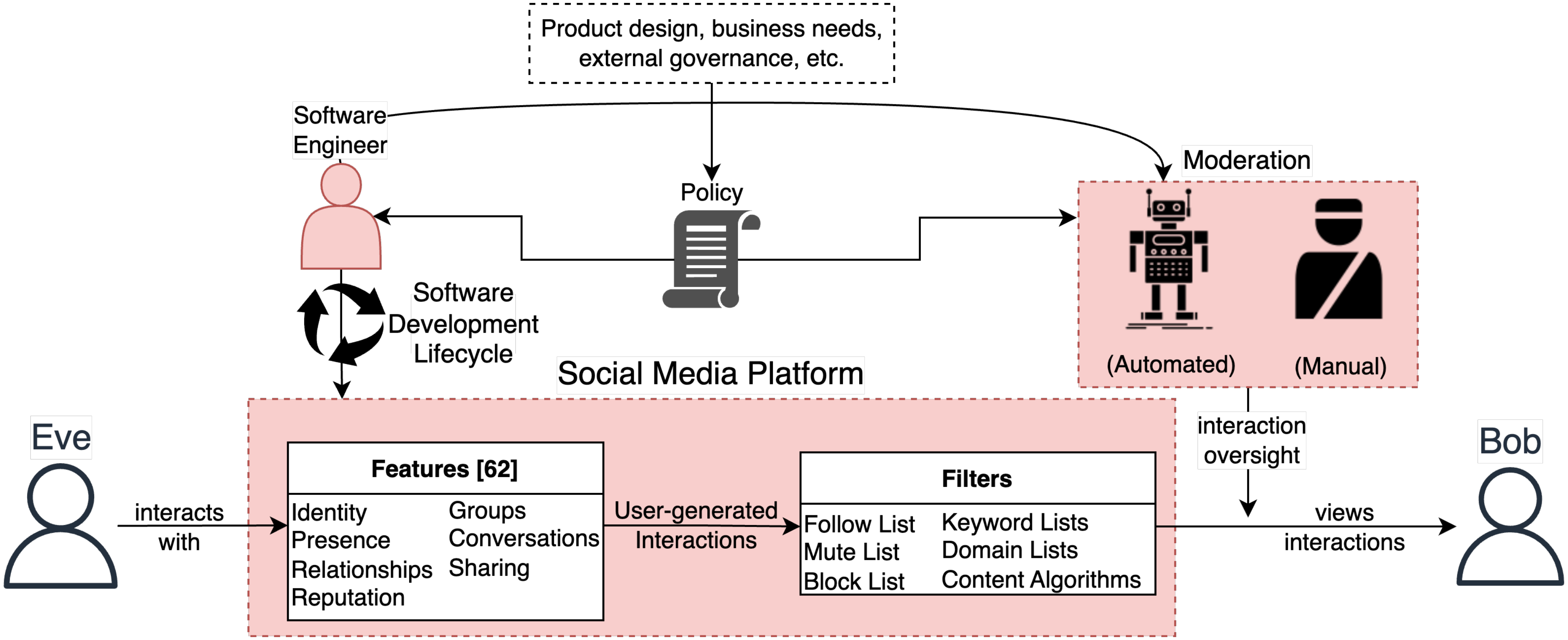
RQ3: What options are proposed in T&S engineering discussions? How are they selected?

T&S Engineering Pattern Taxonomy



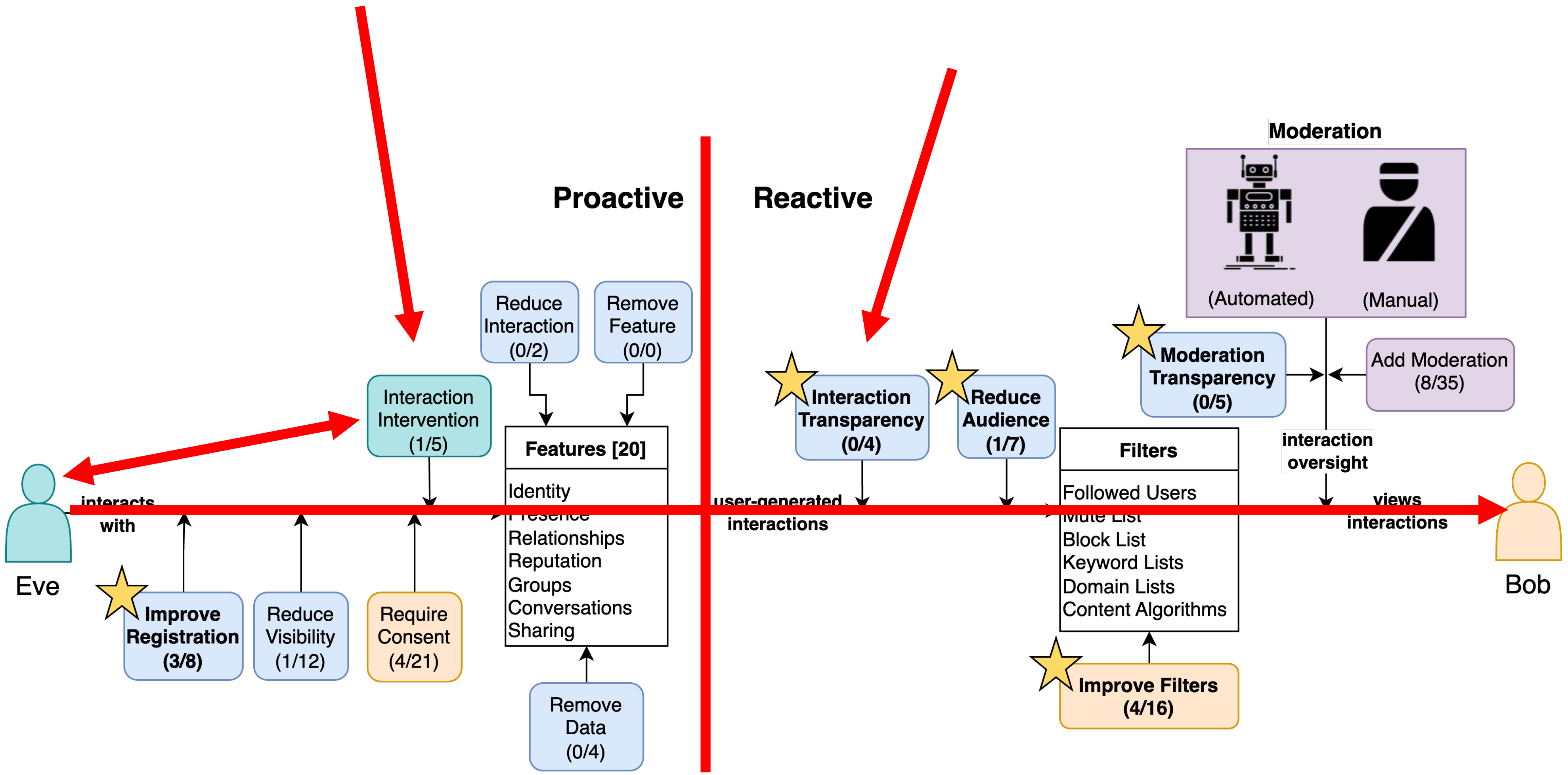
Pattern	**Description**	**Proposed**	**Chosen**
Add moderation	Add or improve moderation tools	20	7
Require consent	Ask for approval from involved stakeholders	15	4
Improve filters	Allow users to better control the content they see	7	3
Reduce visibility	Limit when a feature can be used	8	1
Improve registration	Bolster user trustworthiness checks	6	3
Reduce audience	Limit exposure of content	6	1
Interaction intervention	Intervene before users contact others	3	1
Moderation transparency	Increase clarity of moderation decisions	2	0
Interaction transparency	Clarity of events that occurred between users	3	0
Remove data	Remove unnecessary data from platform	4	0
Reduce interaction	Limit how a feature can be used	2	0
Remove feature	Take out feature	0	0

T&S Engineering Pattern Diagram

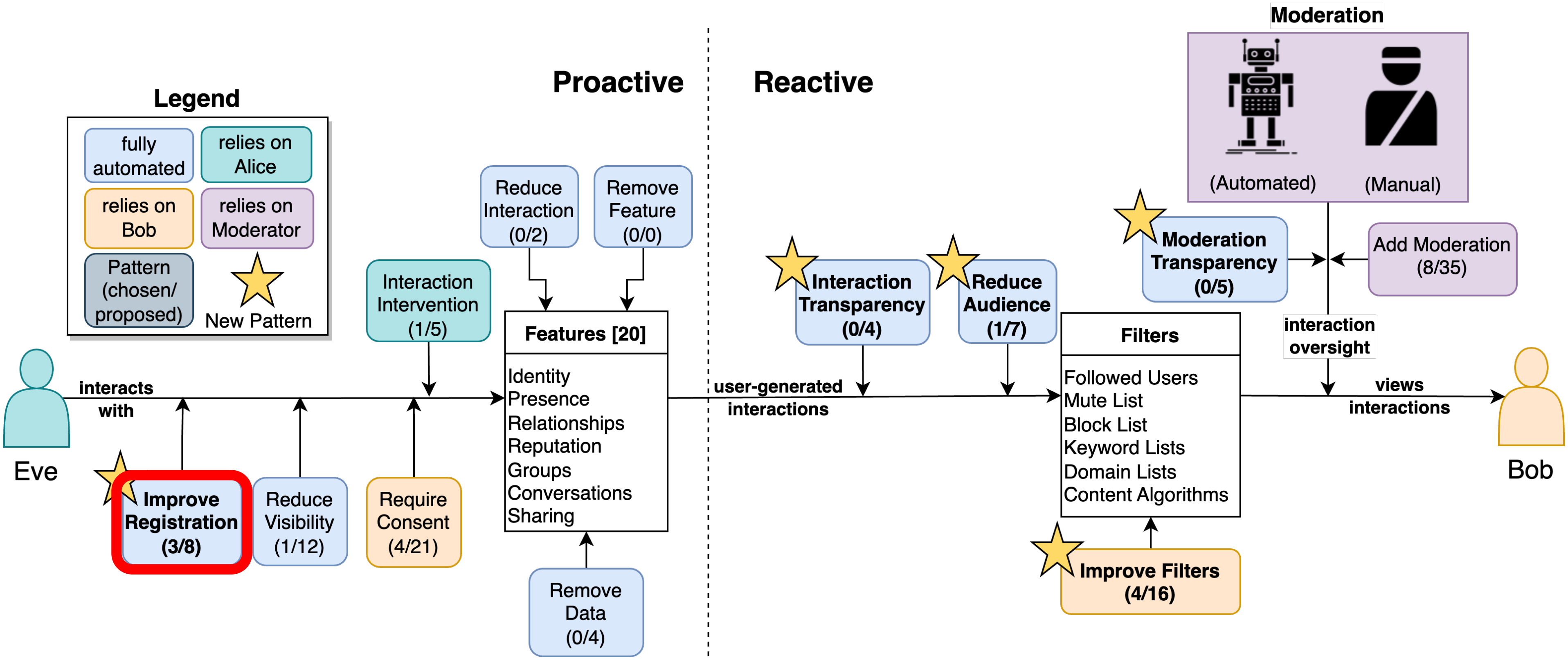


Old context diagram. Will now overlay discovered patterns...

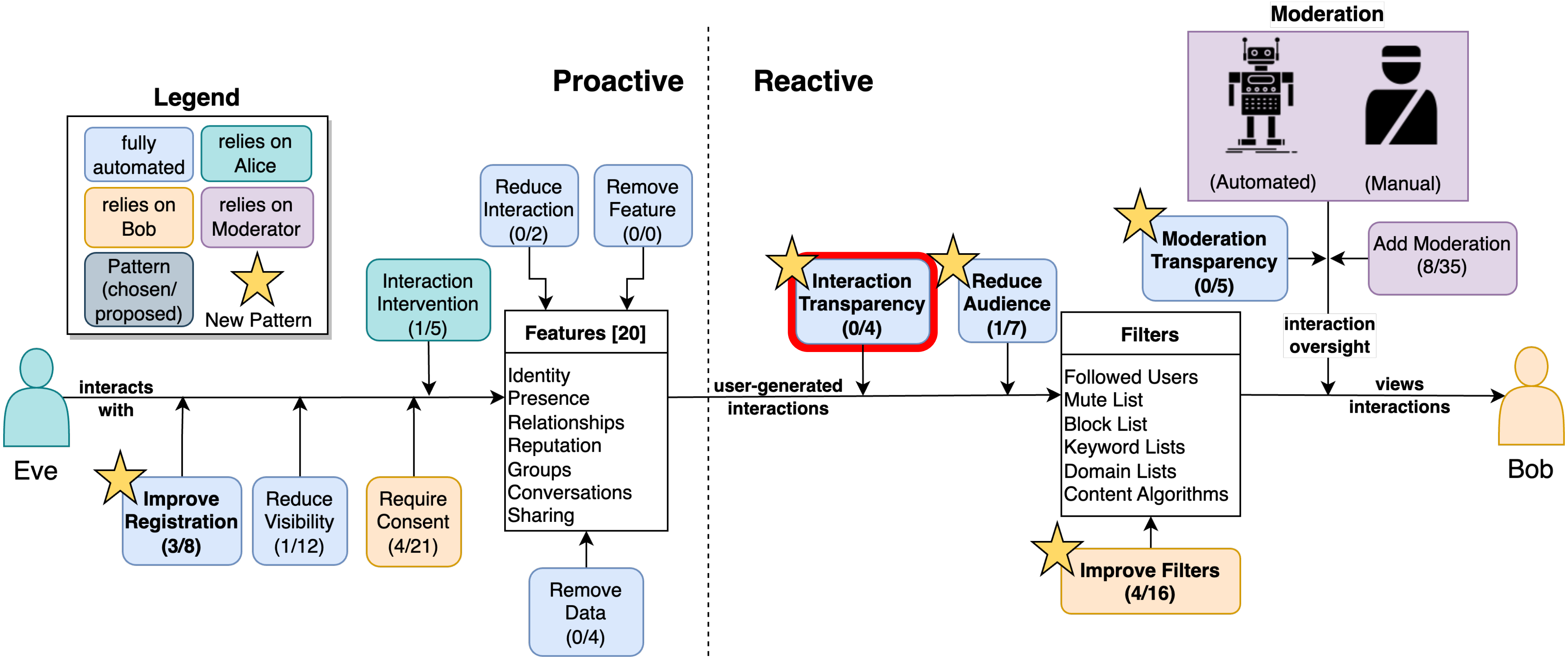
T&S Engineering Pattern Diagram



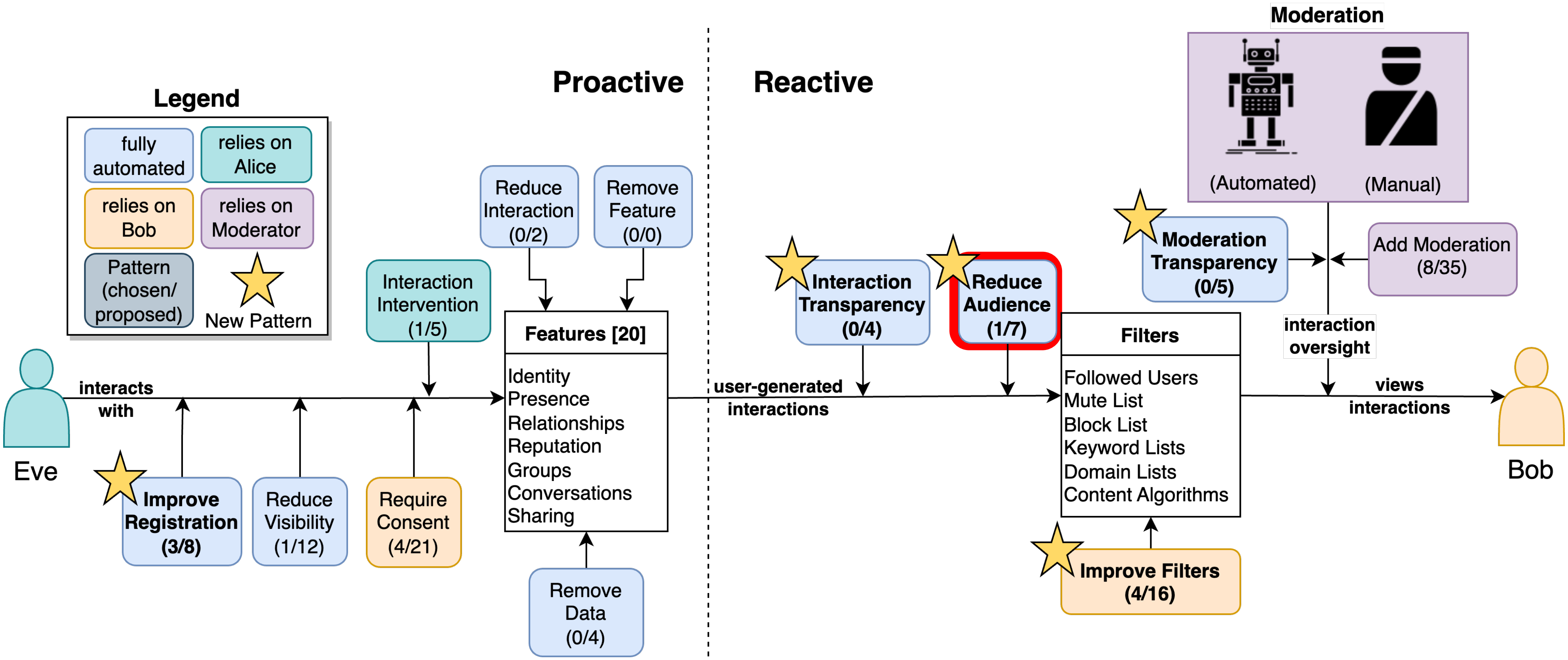
T&S Engineering Pattern Diagram



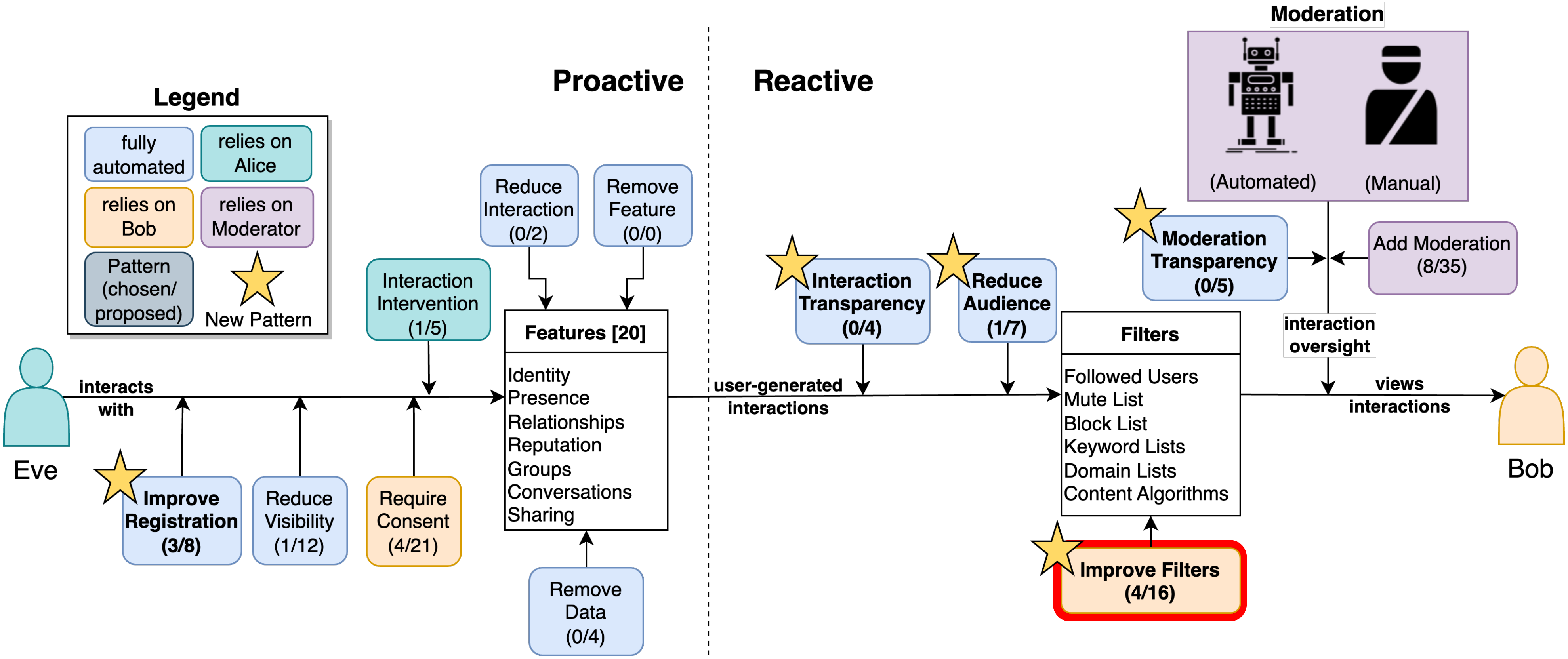
T&S Engineering Pattern Diagram



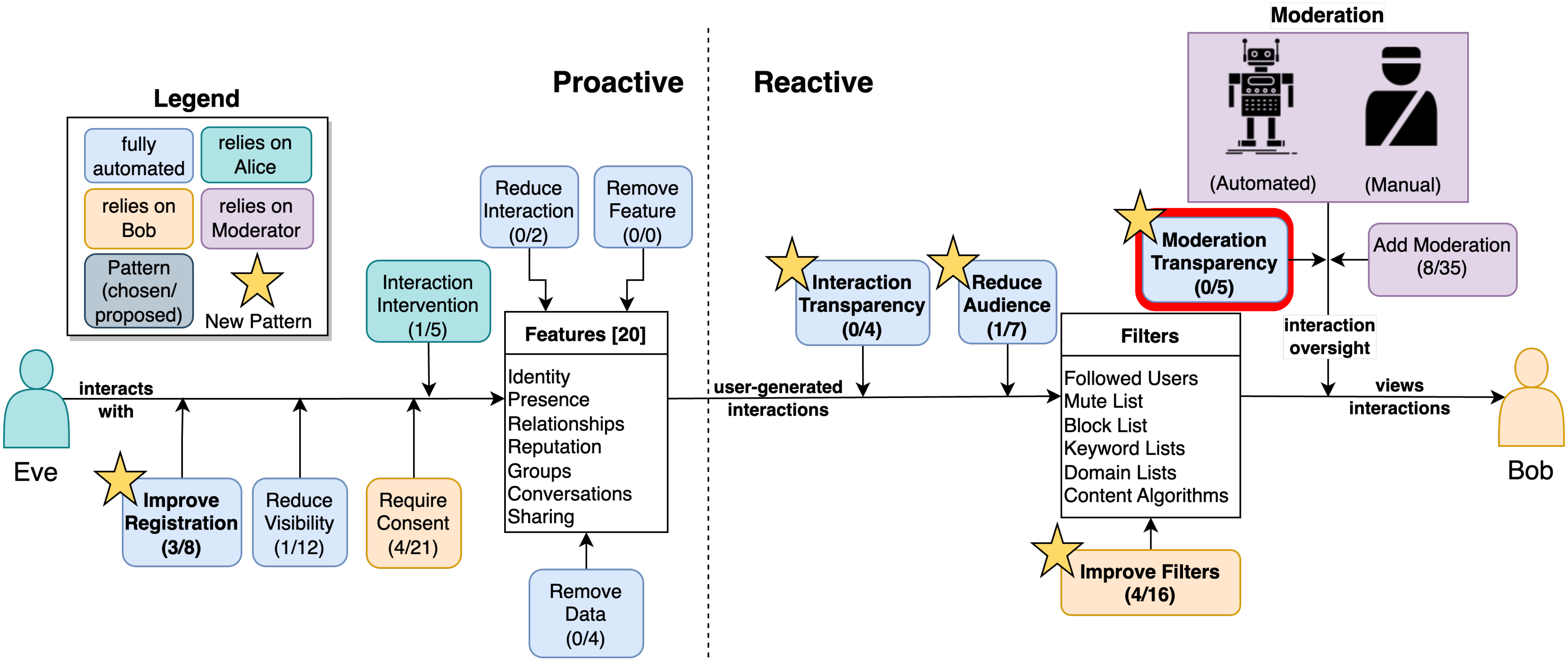
T&S Engineering Pattern Diagram



T&S Engineering Pattern Diagram



T&S Engineering Pattern Diagram



Legend

fully automated

relies on Alice

relies on Bob

relies on Moderator

Pattern (chosen/ proposed)

New Pattern

Proactive

Reactive

Moderation

(Automated)

(Manual)

Eve

interacts with

Improve Registration (3/8)

Reduce Visibility (1/12)

Require Consent (4/21)

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)

user-generated interactions

Interaction Transparency (0/4)

Reduce Audience (1/7)

Filters

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

Improve Filters (4/16)

Moderation Transparency (0/5)

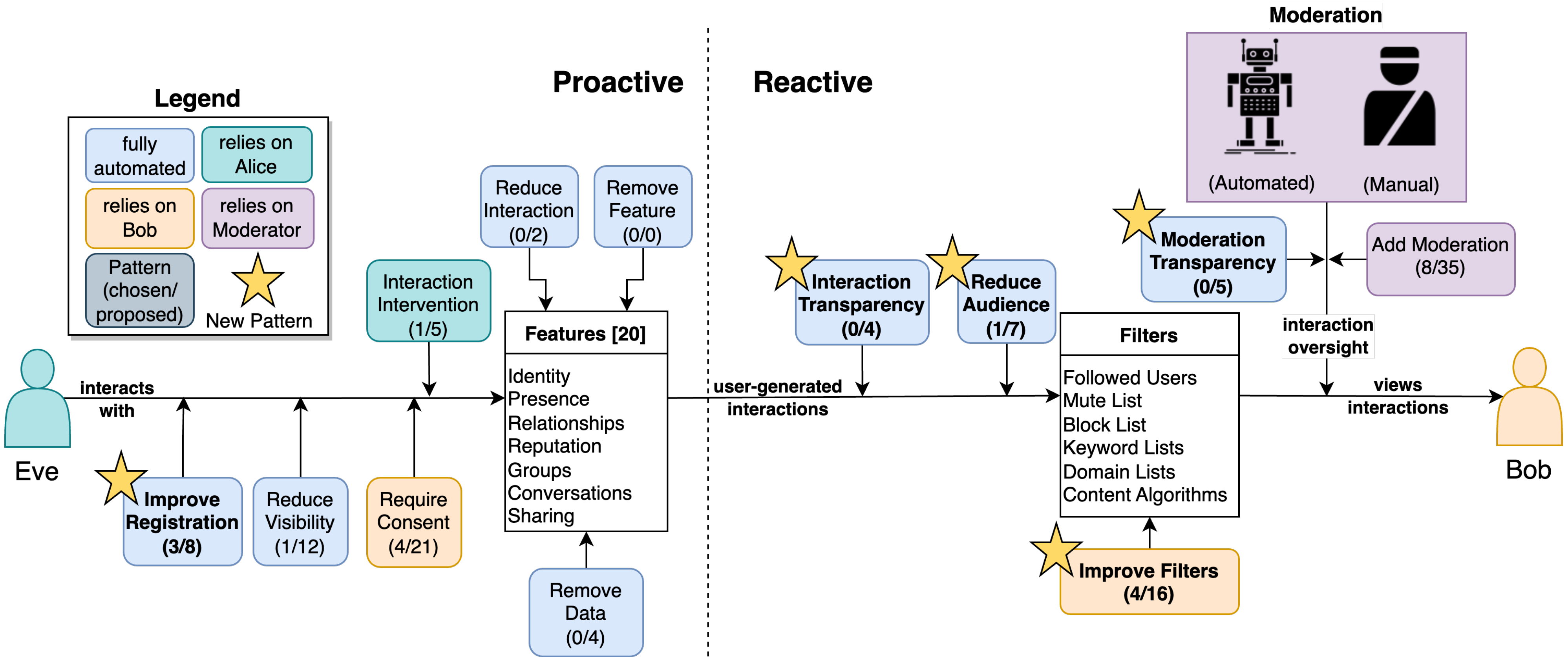
Add Moderation (8/35)

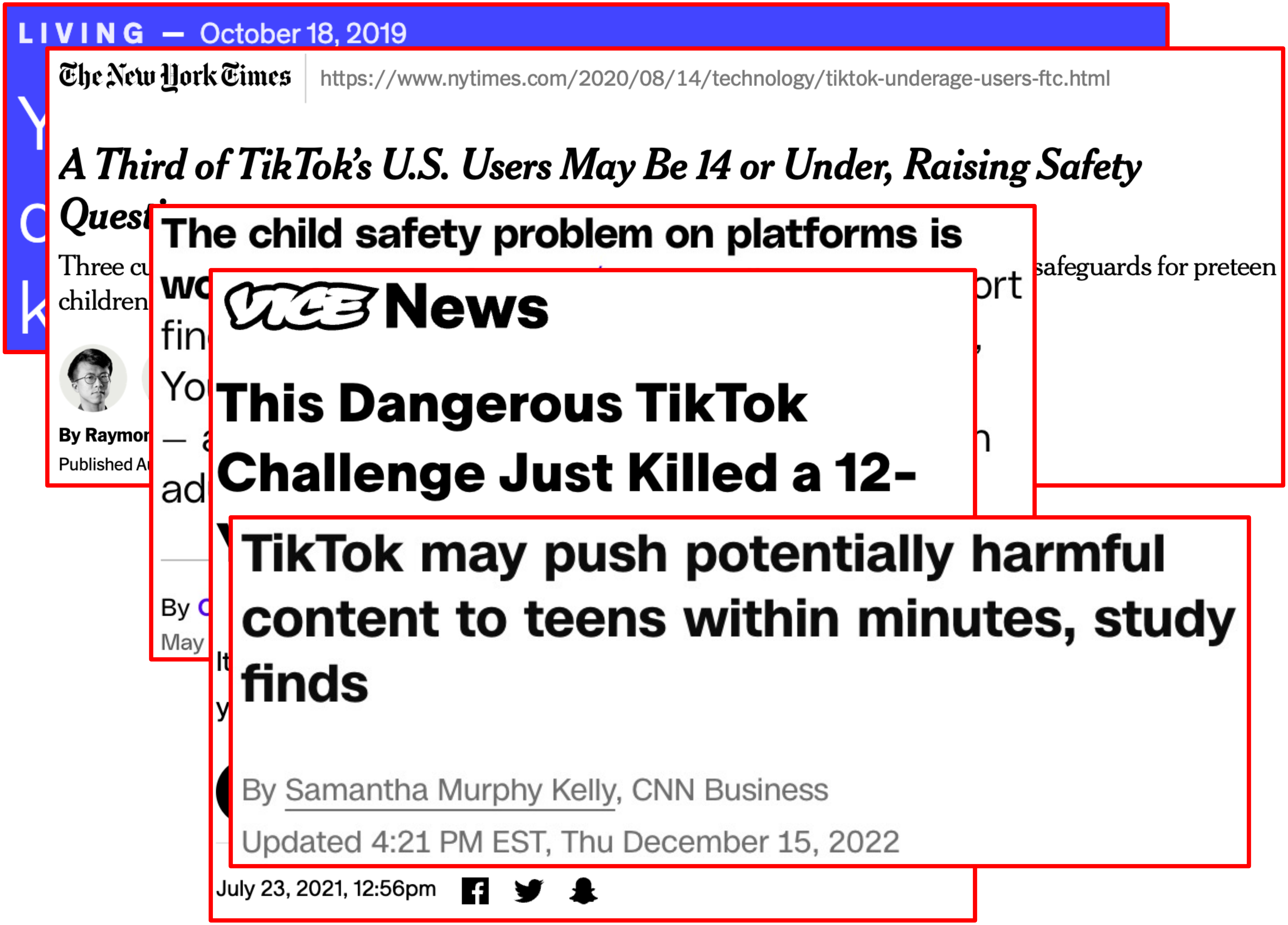
interaction oversight

views interactions

Bob

T&S Engineering Pattern Diagram





LIVING — October 18, 2019

Y

O

k

The New York Times https://www.nytimes.com/2020/08/14/technology/tiktok-underage-users-ftc.html

A Third of TikTok's U.S. Users May Be 14 or Under, Raising Safety Quest

Three cu safeguards for preteen children

By Raymor

Published Au

The child safety problem on platforms is

wo ort

fin ,

Yo

— a n

ad

By C

May

VICE News

This Dangerous TikTok Challenge Just Killed a 12-

Y

It

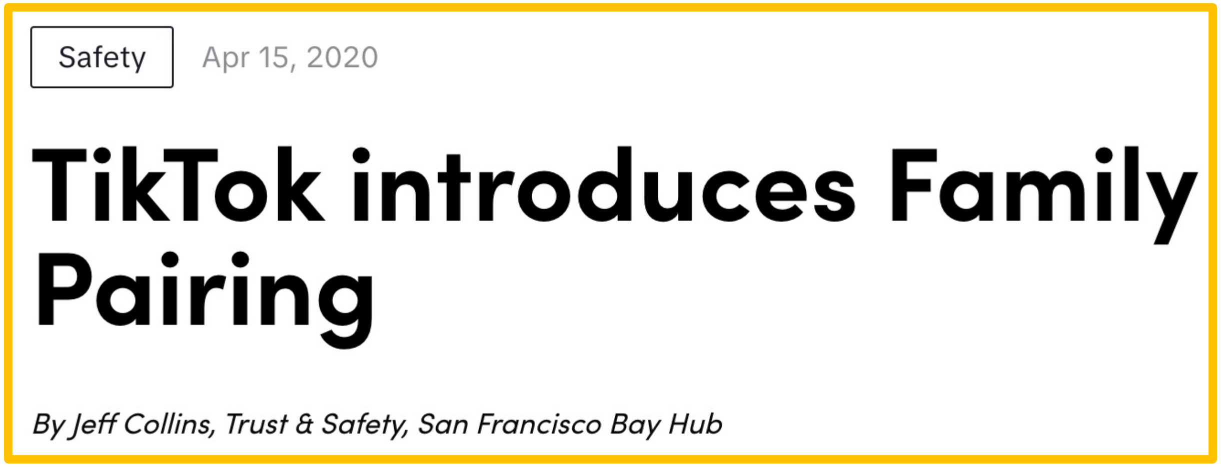
y

July 23, 2021, 12:56pm

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



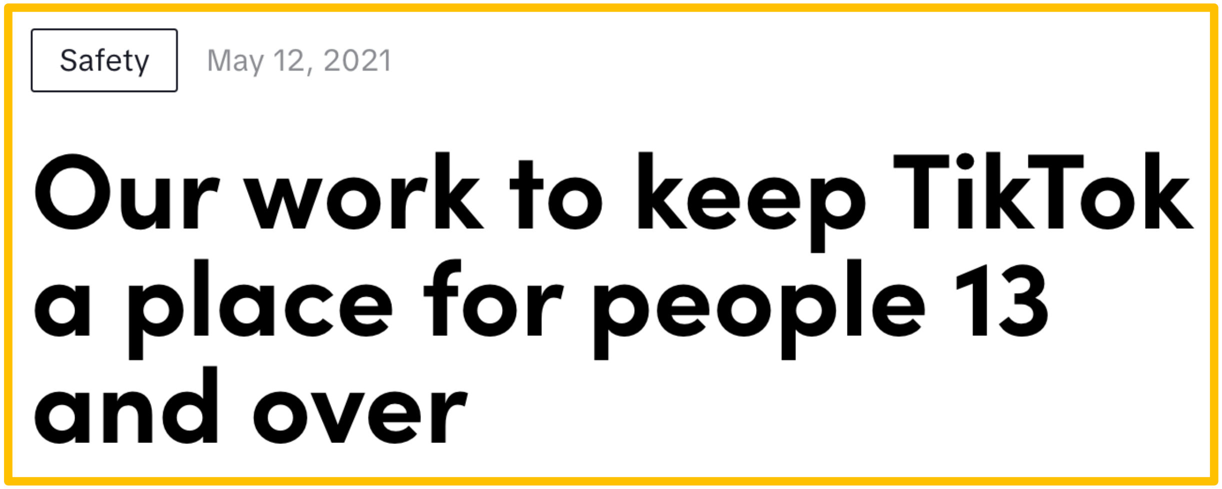
Safety

Apr 15, 2020

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

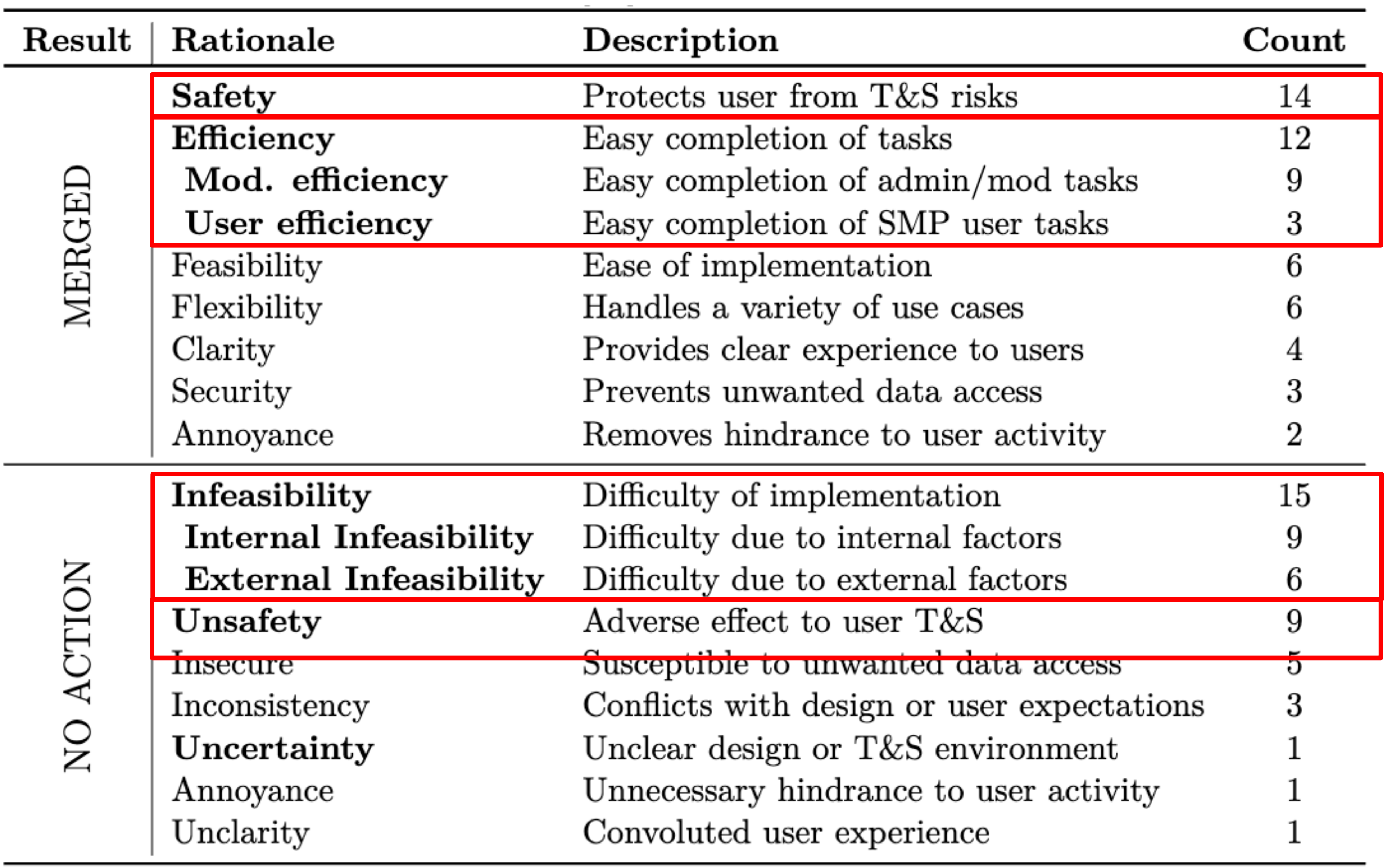


Safety May 12, 2021

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

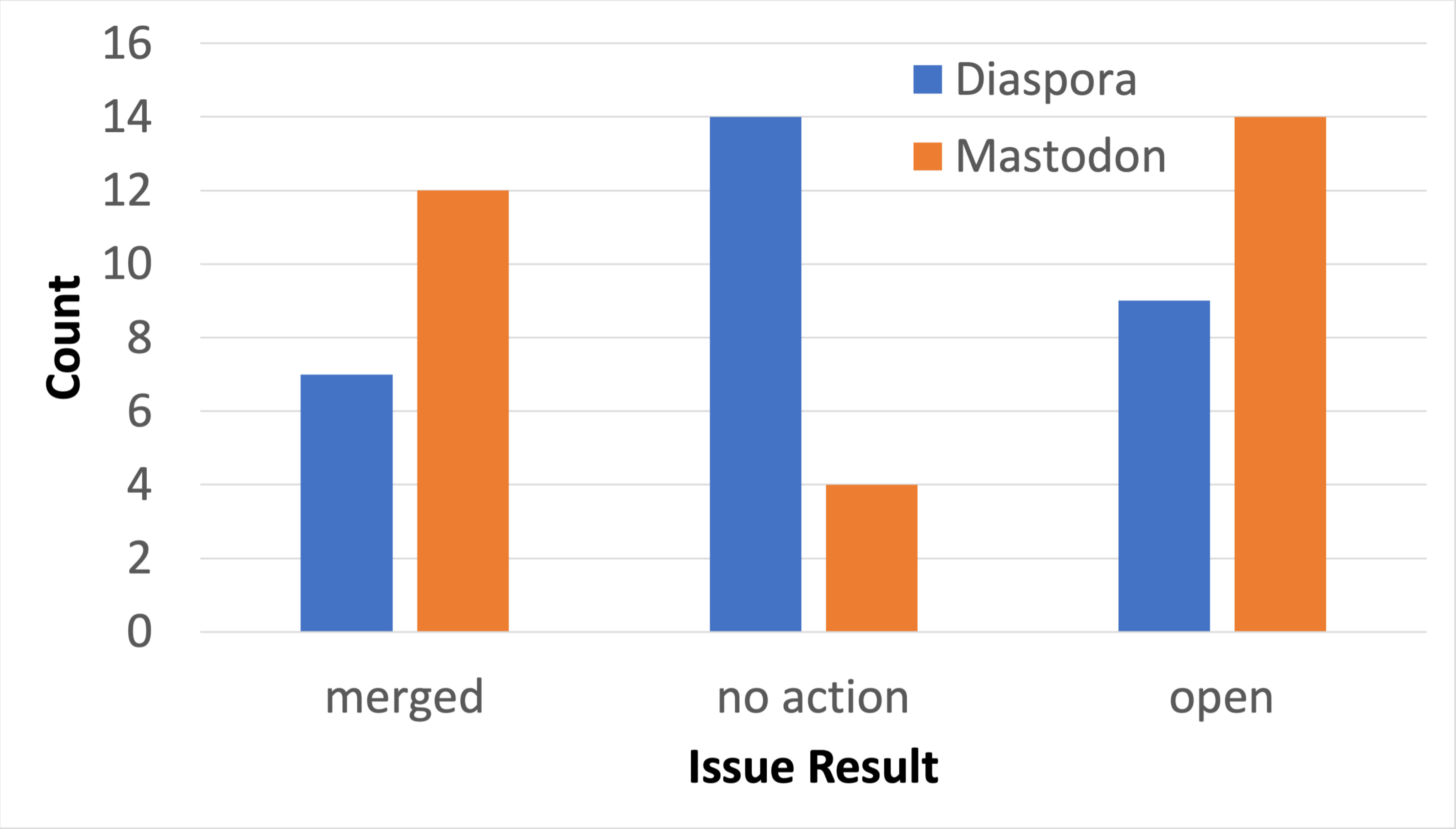
- Prevent underage users from signing up -> improve verification
- Remove underage accounts -> add moderation
- Bring transparency to our actions -> moderation transparency
- Age-appropriate environment for new young users
  - Remove access to LIVE -> reduce visibility
  - Remove access to DMs -> reduce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

Rationale Taxonomy



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

Diaspora is more passive. 40% of issues are open.



RQ3: Key Findings

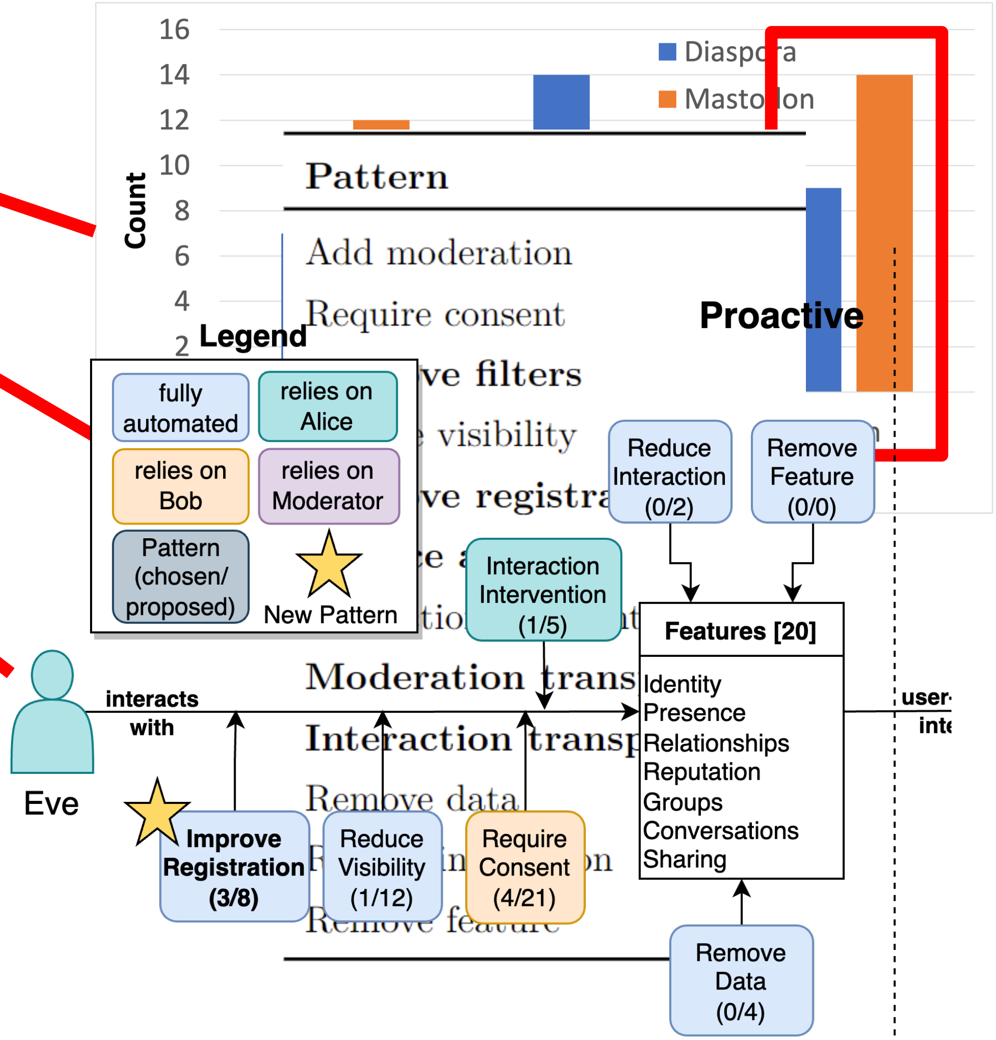
- Most commonly proposed and chosen pattern is *add moderation*
- Reactive patterns are chosen more often
- New patterns are primarily reactive
- 38% of identified T&S issues remain open
- T&S issue resolution is slow

Outline

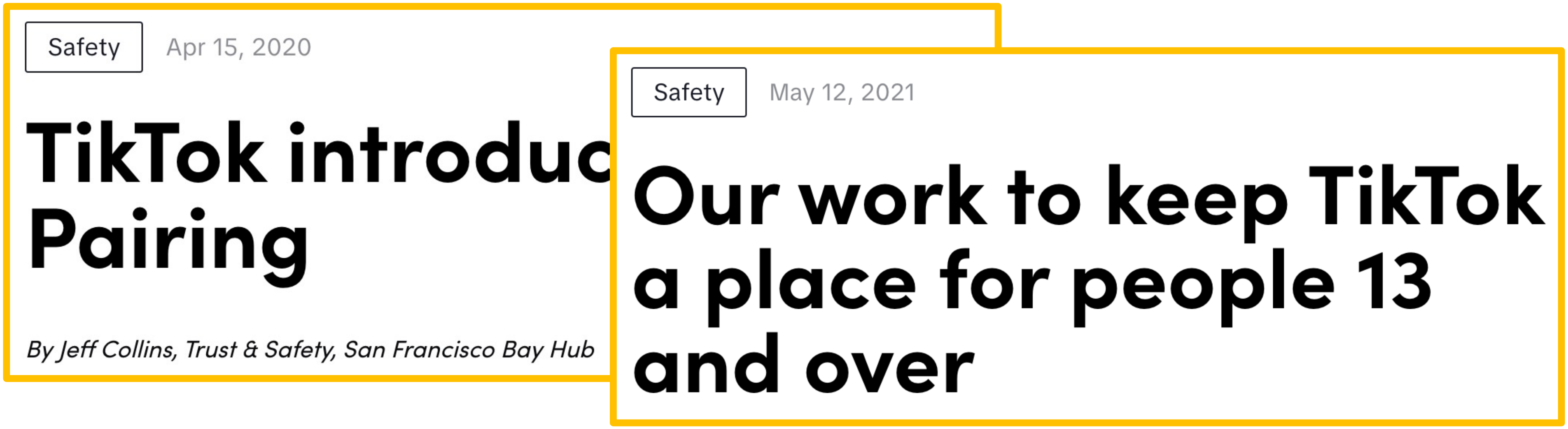
- Background
- Research Questions
- Methodology
- Results
- Discussion
- Threats to Validity

Recommendations for OSS SMPs

- Communicate existing risks
- Document risks and treatments
- Explore proactive solutions



What other patterns could be used?



Safety Apr 15, 2020

TikTok introduc
Pairing

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

Safety May 12, 2021

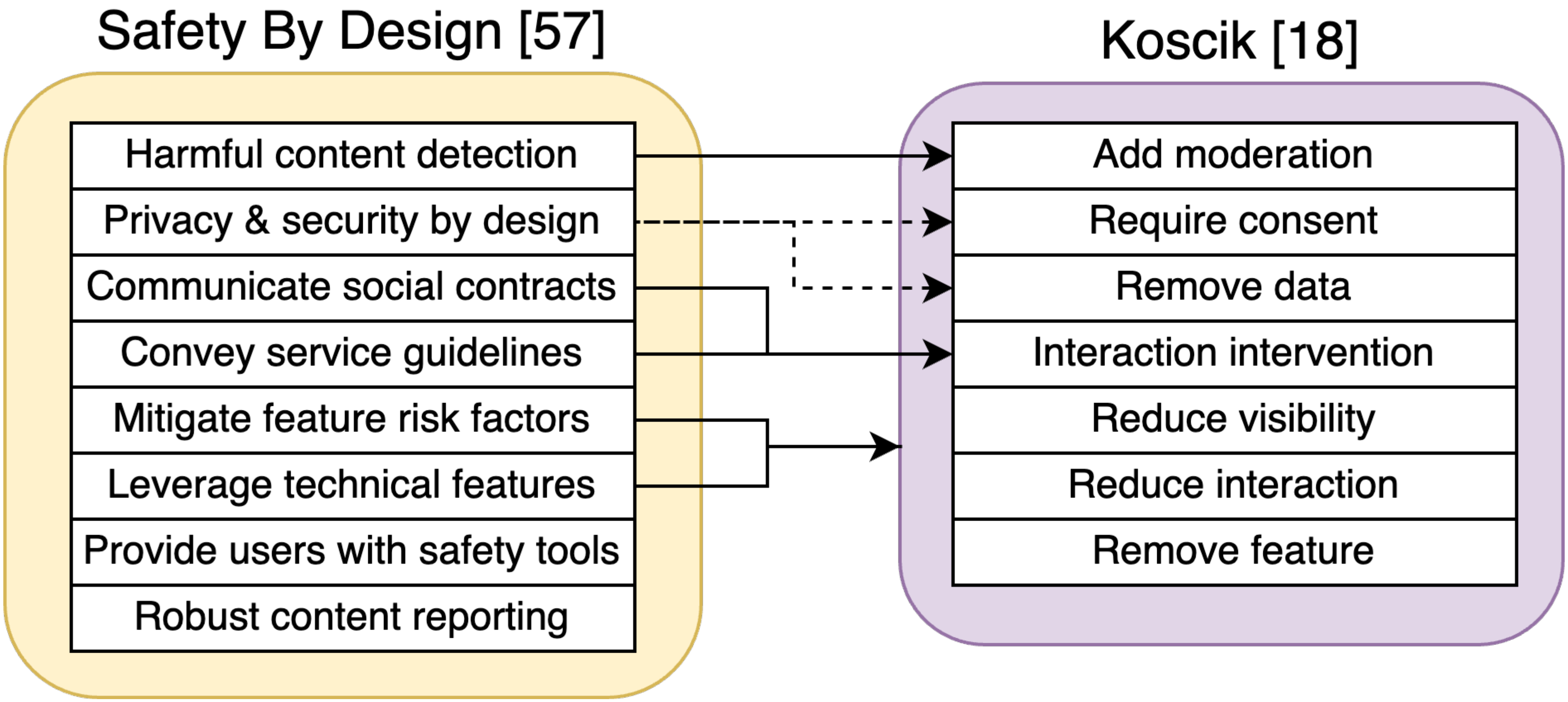
Our work to keep TikTok a place for people 13 and over

- 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

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?



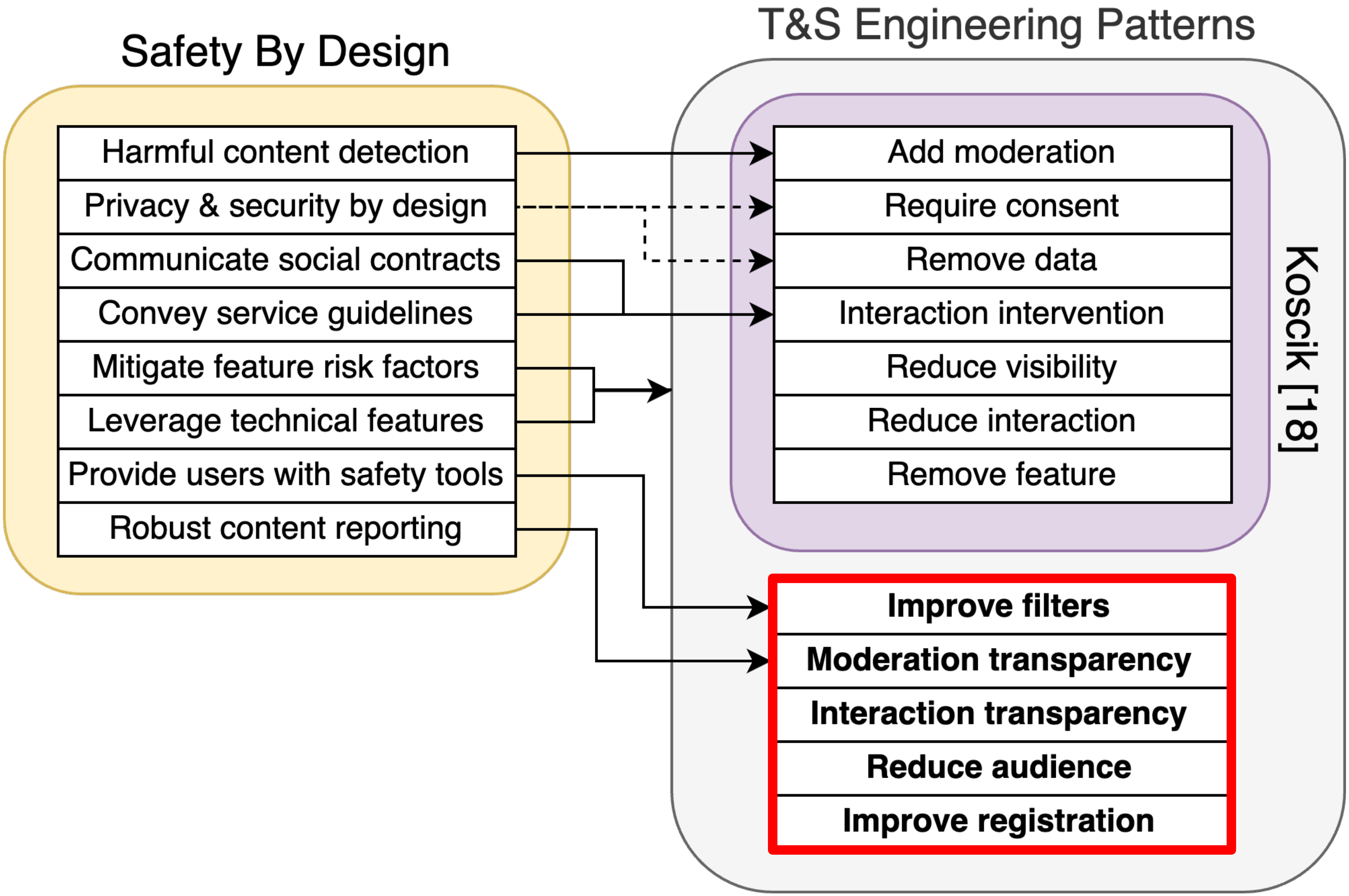
Safety By Design [57]

- 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

Koscik [18]

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

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

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

Smith, G. (2007, April 7). *Social Software Building Blocks*. https://web.archive.org/web/20171123070545/http://nform.com/ideas/social-software-building-blocks

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

DataReportal. (2023, January). *Global Social Media Statistics*. https://datareportal.com/social-media-users

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

Cryst, E., Grossman, S., Hancock, J., Stamos, A., & Thiel, D. (2021). Introducing the Journal of Online Trust and Safety. *Journal of Online Trust and Safety*, *1*(1), Article 1. https://tsjournal.org/index.php/jots/article/view/8

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

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

Seering, J., Fang, T., Damasco, L., Chen, M. "Cherie," Sun, L., & Kaufman, G. (2019). Designing User Interface Elements to Improve the Quality and Civility of Discourse in Online Commenting Behaviors. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3290605.3300836

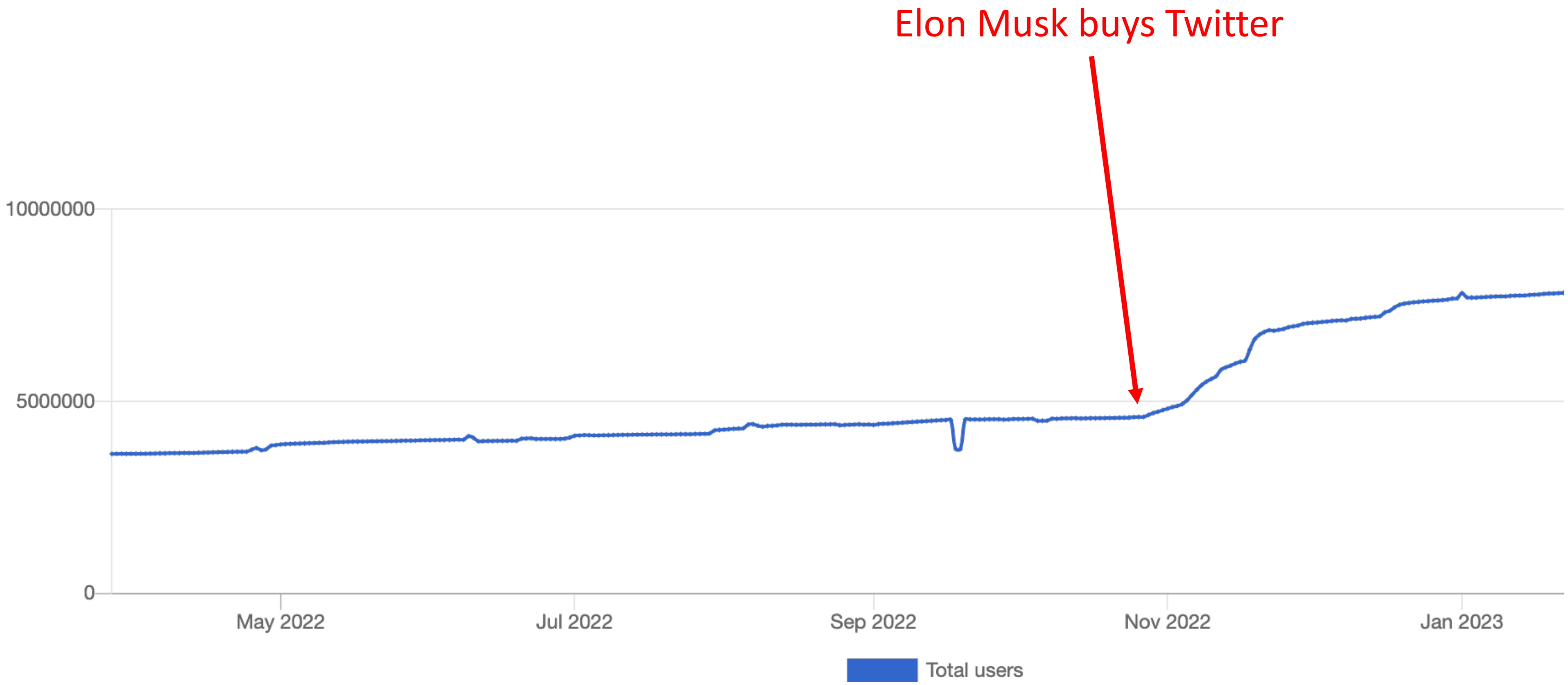
eSafety Commissioner. (n.d.). *Safety by Design*. ESafety Commissioner. https://web.archive.org/web/20220308081249/https://www.esafety.gov.au/industry/safety-by-design

Ko, A. J., & Chilana, P. K. (2011). Design, discussion, and dissent in open bug reports. *Proceedings of the 2011 IConference*, 106–113. https://doi.org/10.1145/1940761.1940776

Cavoukian, A., & Dixon, M. (2013). Privacy and security by design: An enterprise architecture approach. Information and Privacy Commissioner of Ontario, Canada.

Thank You!

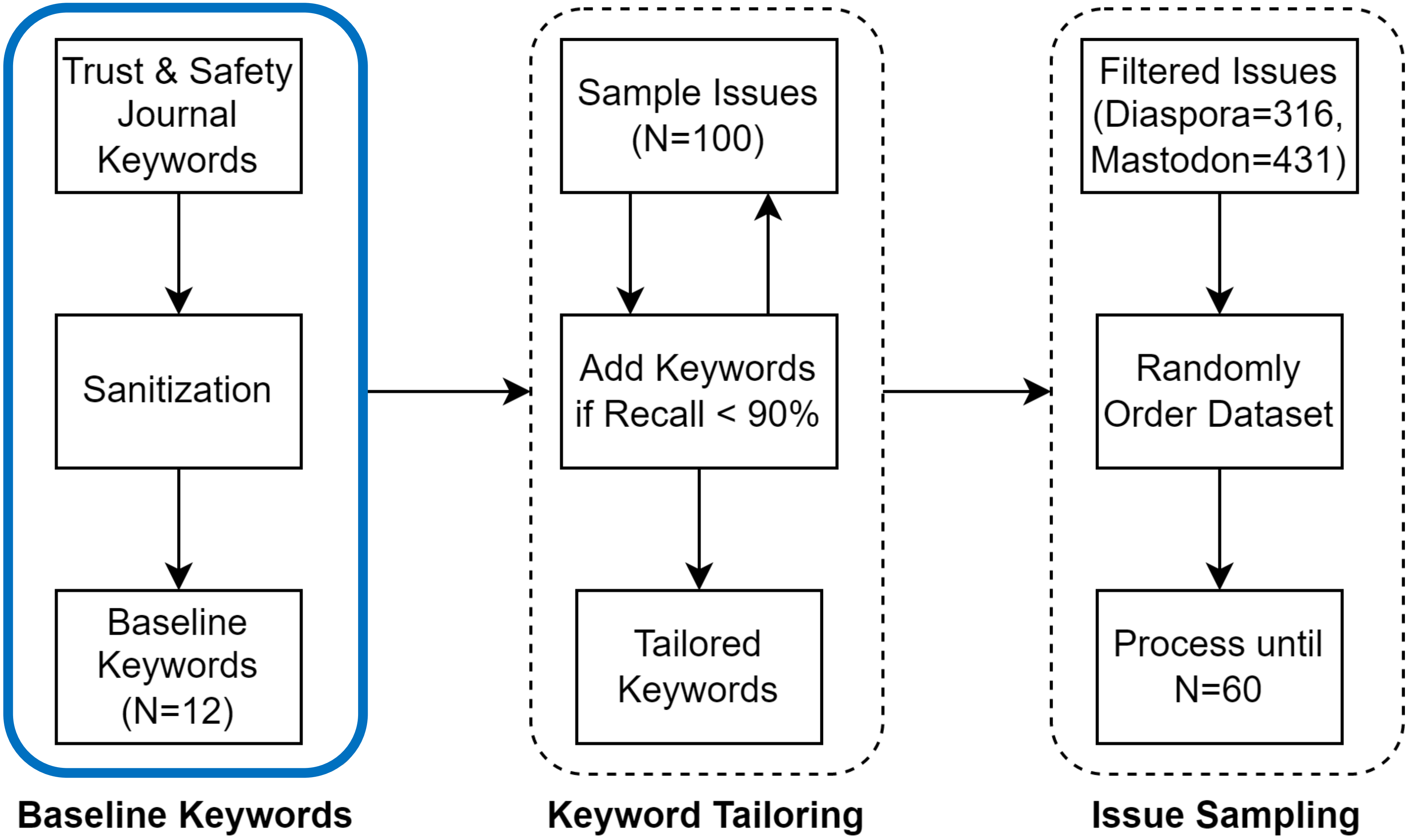
Bonus Slides



Elon Musk buys Twitter

Mastodon accounts over time []

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

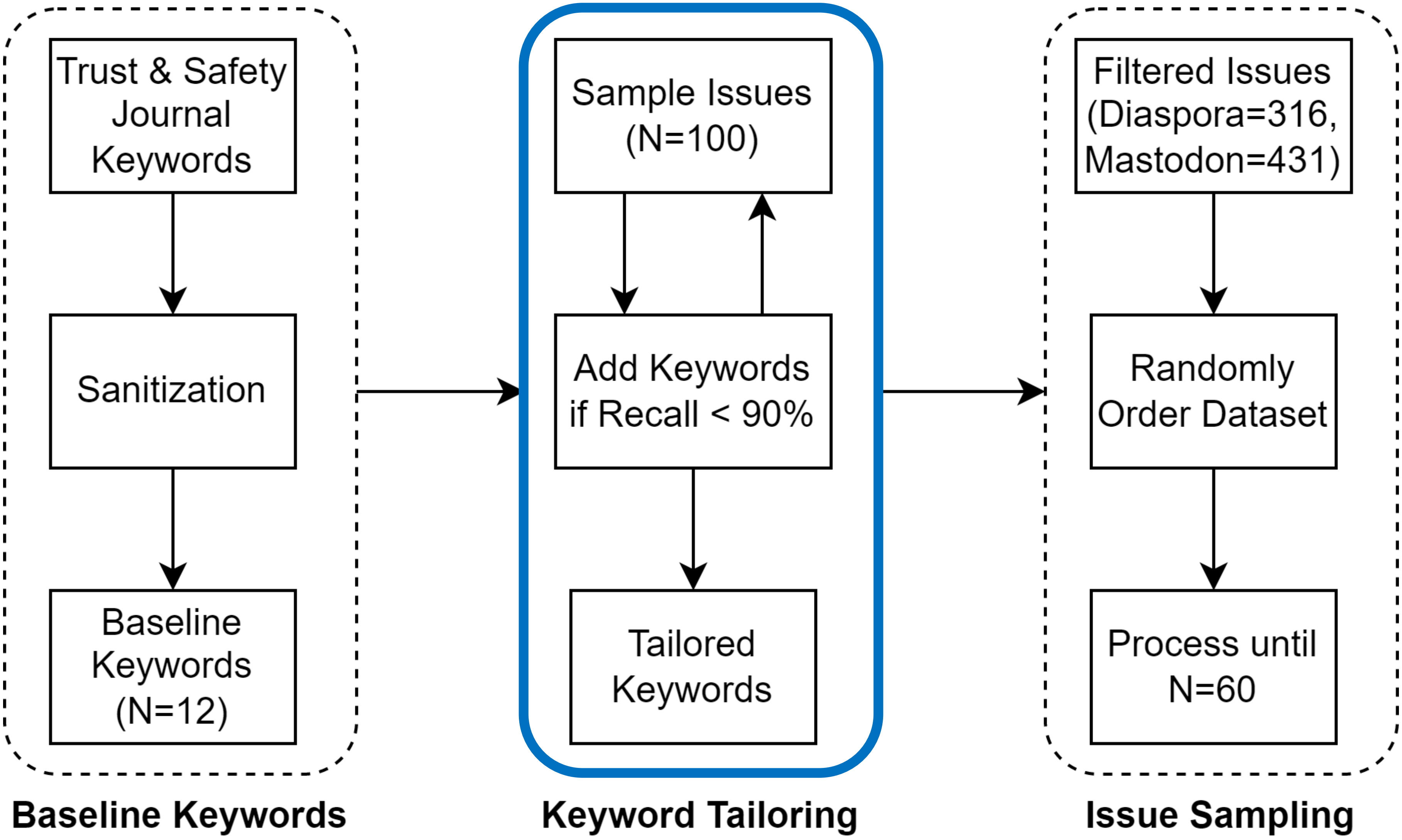
Process until N=60

Baseline Keywords

Keyword Tailoring

Issue Sampling

Issue Selection



Trust & Safety Journal Keywords

Sanitization

Baseline Keywords (N=12)

Baseline Keywords

Sample Issues (N=100)

Add Keywords if Recall < 90%

Tailored Keywords

Keyword Tailoring

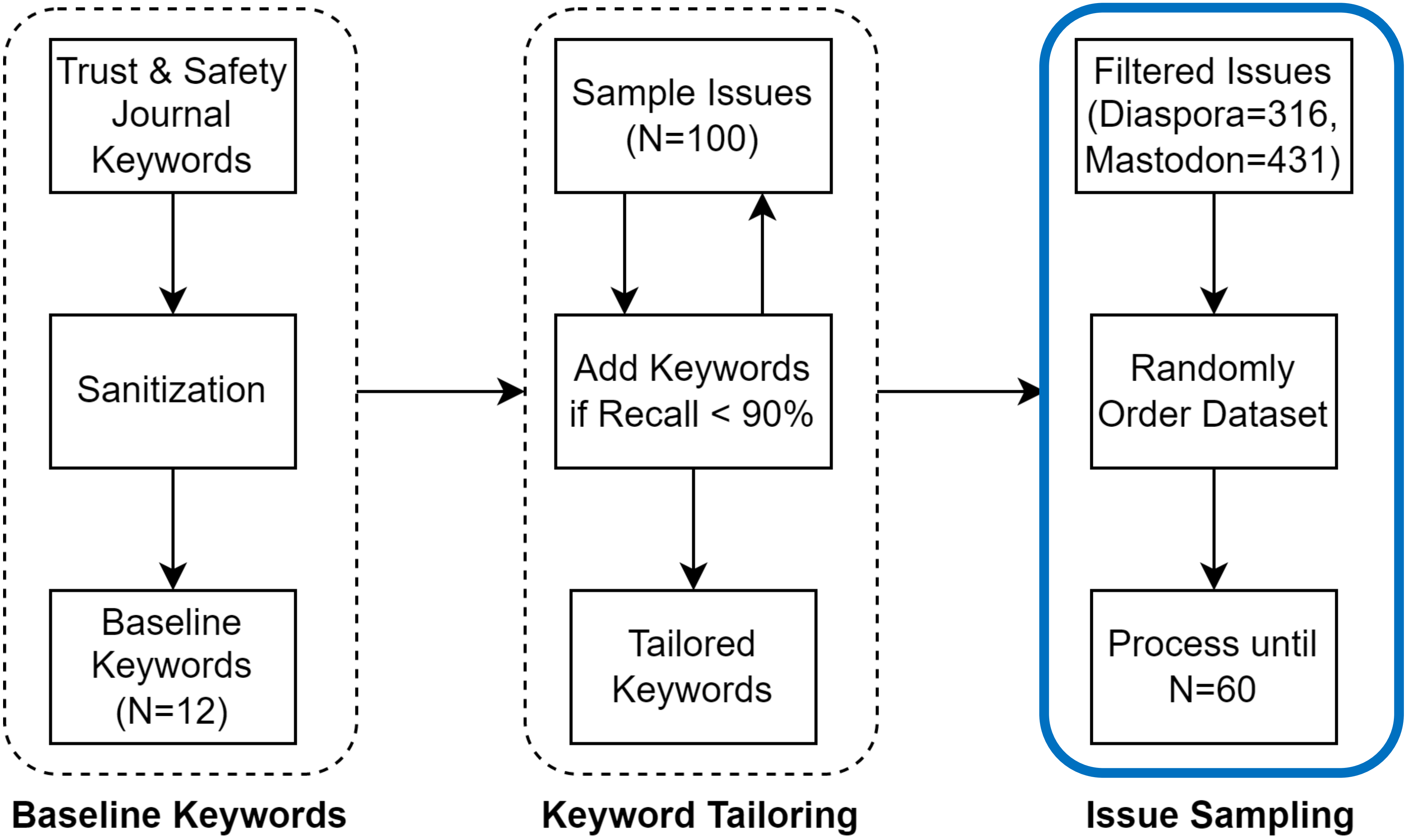
Filtered Issues (Diaspora=316, Mastodon=431)

Randomly Order Dataset

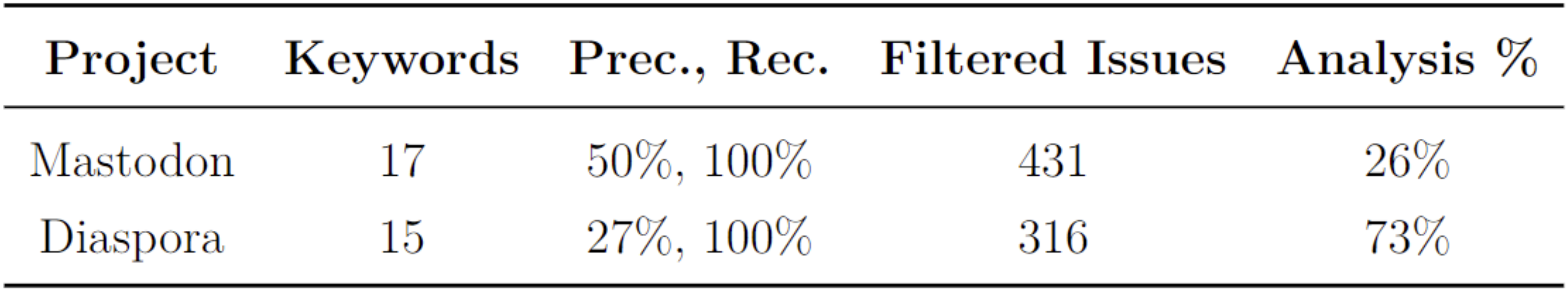
Process until N=60

Issue Sampling

Issue Selection



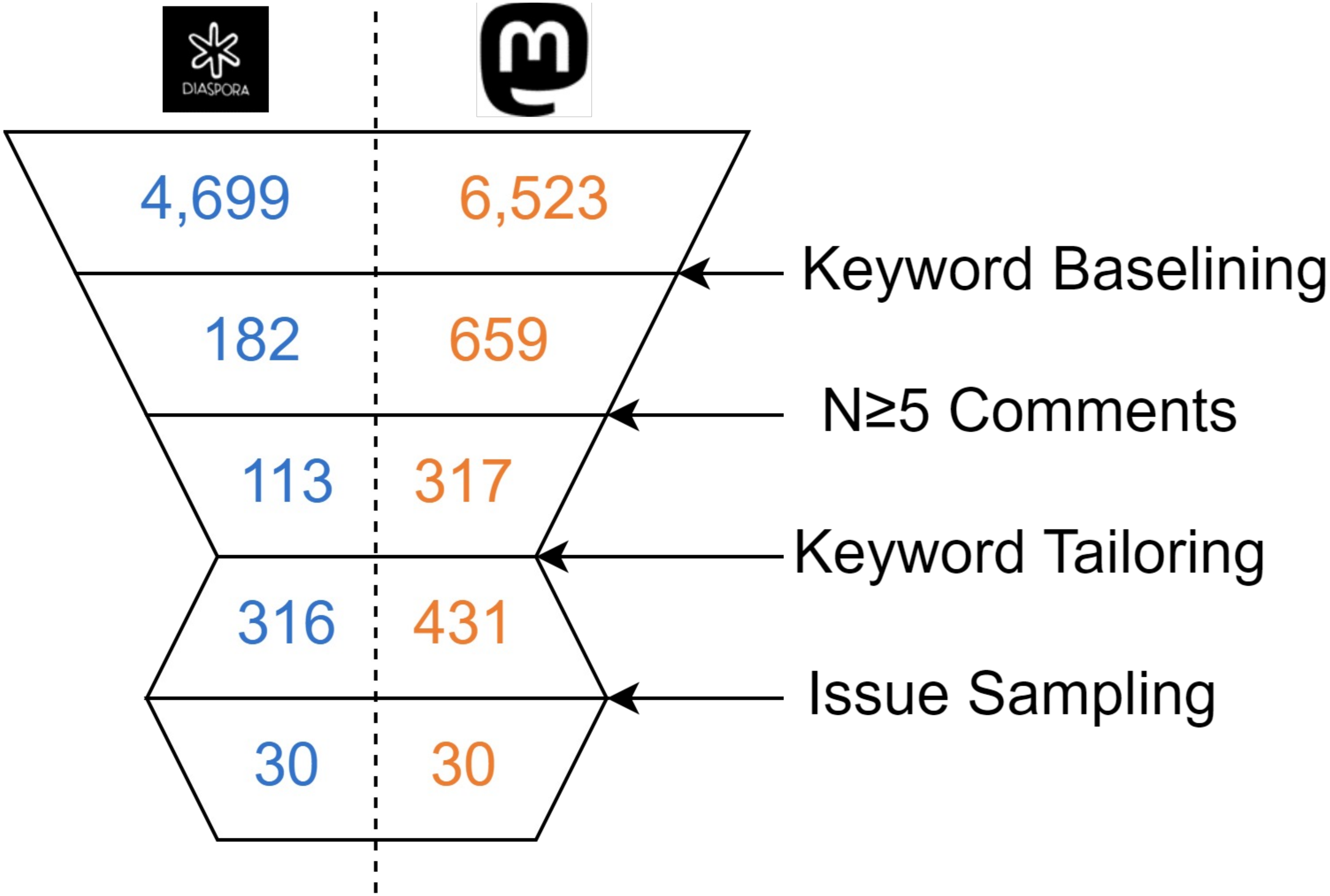
Issue Selection – Keyword Results



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

SMP filtering results

Issue Selection – Eligible Population



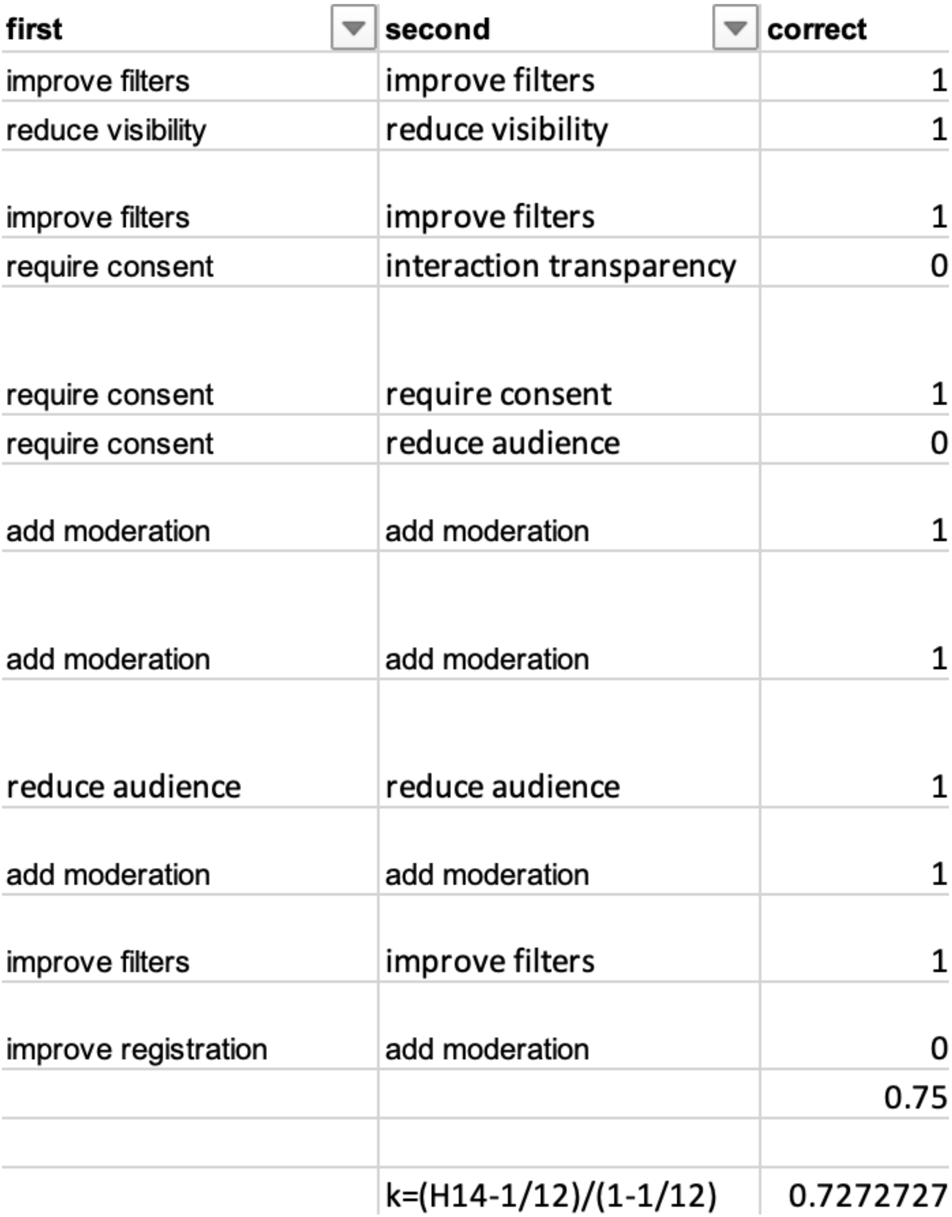
Issue Analysis – Issue Data Mapping

GitHub Data	Use in study
Issue Title	Used for discussion modeling (if needed).
Issue Status (open/closed)	Used for *issue result* ("open", "merged", "no action")
Issue open date	Used for longitudinal analysis.
Issue closure date	Used for longitudinal analysis
Issue comments	Separated comments into sentences. Used to determine *issue type* ("bug" or "feature request"). Used to label SMP feature. Coded relevant sentences as *risk, option, rationale,* and *chosen.*
Linked pull requests	Used to distinguish between "merged" and "no action" *issue result.* Used to determine which *options* to mark as *chosen.*

Inter-rater Agreement – Kappa Calculation

$$\kappa \equiv \frac{p_o - p_e}{1 - p_e}$$

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



first	second	correct
improve filters	improve filters	1
reduce visibility	reduce visibility	1
improve filters	improve filters	1
require consent	interaction transparency	0
require consent	require consent	1
require consent	reduce audience	0
add moderation	add moderation	1
add moderation	add moderation	1
reduce audience	reduce audience	1
add moderation	add moderation	1
improve filters	improve filters	1
improve registration	add moderation	0
		0.75
	k=(H14-1/12)/(1-1/12)	0.7272727

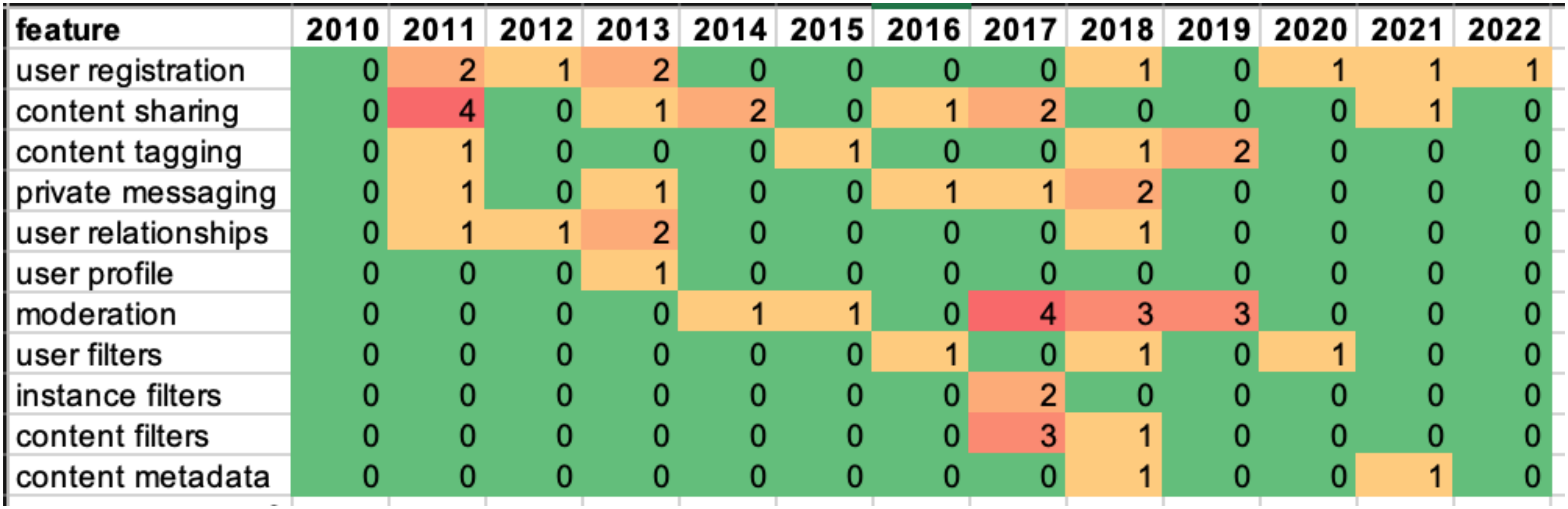
Issue Results

result	count	frequency
merged	19	0.31666667
no action	18	0.3
open	23	0.38333333

Features – Diaspora vs. Mastodon

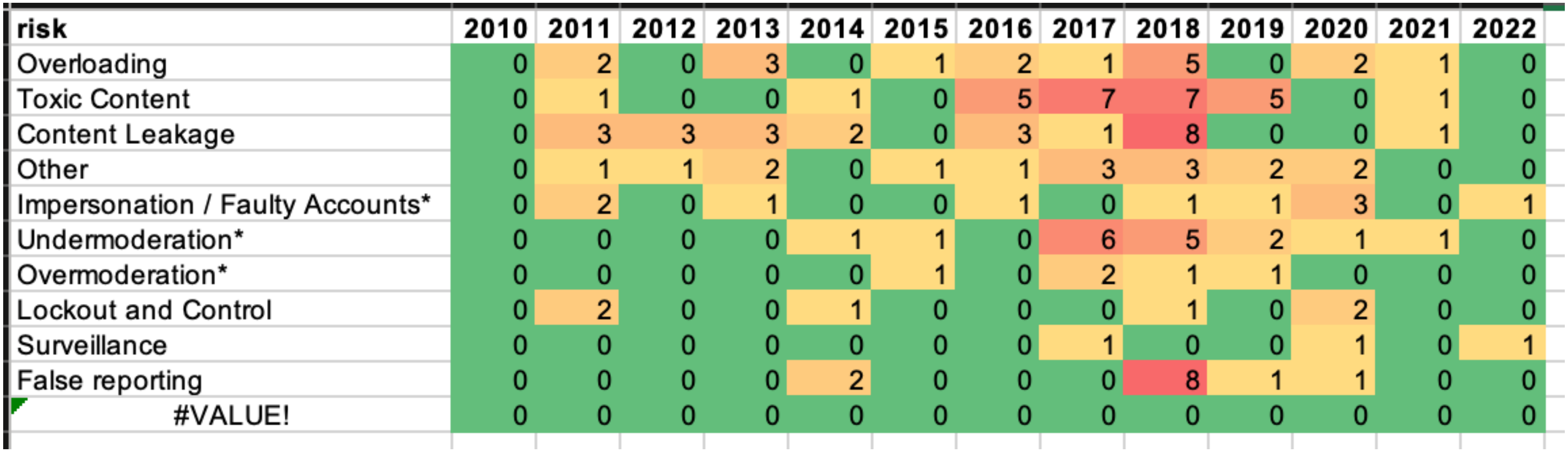
feature	diaspora	mastodon
user registration	6	3
content sharing	9	2
content tagging	3	2
private messaging	3	3
user relationships	4	1
user profile	1	0
moderation	4	8
user filters	0	3
instance filters	0	2
content filters	0	4
content metadata	0	2

Features Over Time



feature	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
user registration	0	2	1	2	0	0	0	0	1	0	1	1	1
content sharing	0	4	0	1	2	0	1	2	0	0	0	1	0
content tagging	0	1	0	0	0	1	0	0	1	2	0	0	0
private messaging	0	1	0	1	0	0	1	1	2	0	0	0	0
user relationships	0	1	1	2	0	0	0	0	1	0	0	0	0
user profile	0	0	0	1	0	0	0	0	0	0	0	0	0
moderation	0	0	0	0	1	1	0	4	3	3	0	0	0
user filters	0	0	0	0	0	0	1	0	1	0	1	0	0
instance filters	0	0	0	0	0	0	0	2	0	0	0	0	0
content filters	0	0	0	0	0	0	0	3	1	0	0	0	0
content metadata	0	0	0	0	0	0	0	0	1	0	0	1	0

Risks Over Time



risk	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
Overloading	0	2	0	3	0	1	2	1	5	0	2	1	0
Toxic Content	0	1	0	0	1	0	5	7	7	5	0	1	0
Content Leakage	0	3	3	3	2	0	3	1	8	0	0	1	0
Other	0	1	1	2	0	1	1	3	3	2	2	0	0
Impersonation / Faulty Accounts*	0	2	0	1	0	0	1	0	1	1	3	0	1
Undermoderation*	0	0	0	0	1	1	0	6	5	2	1	1	0
Overmoderation*	0	0	0	0	0	1	0	2	1	1	0	0	0
Lockout and Control	0	2	0	0	1	0	0	0	1	0	2	0	0
Surveillance	0	0	0	0	0	0	0	1	0	0	1	0	1
False reporting	0	0	0	0	2	0	0	0	8	1	1	0	0
#VALUE!	0	0	0	0	0	0	0	0	0	0	0	0	0

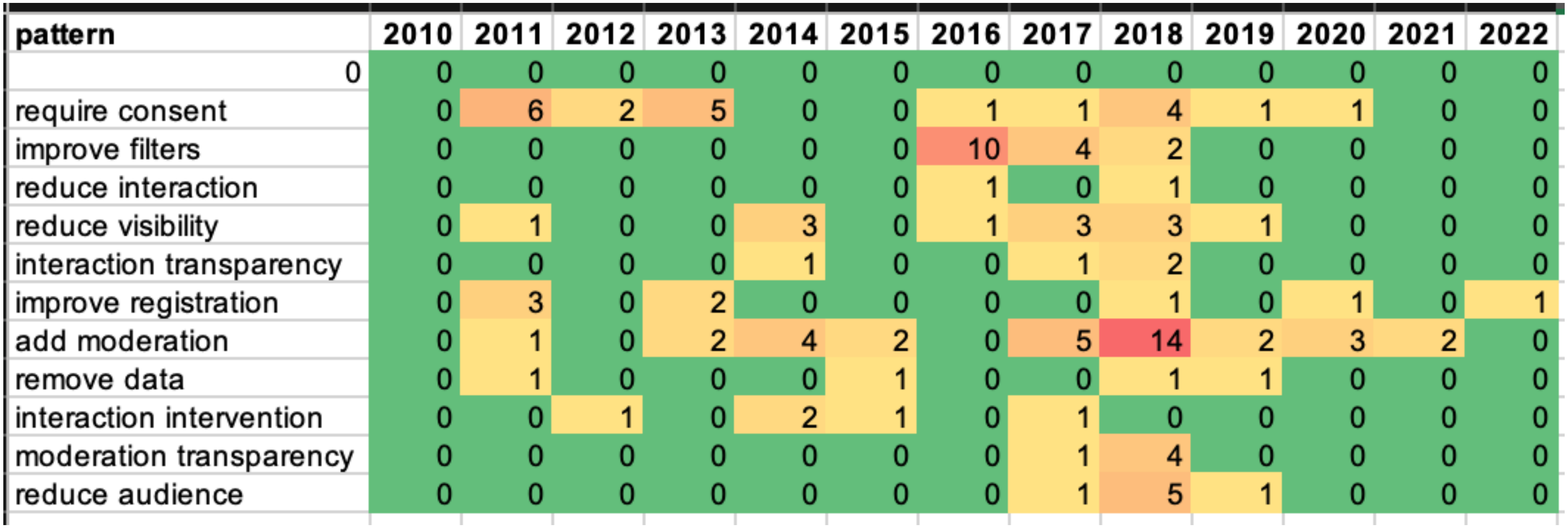
Risks – Diaspora vs. Mastodon

risk	diaspora	mastodon
Overloading	6	11
Toxic Content	5	22
Content Leakage	19	5
Other	5	11
Impersonation / Faulty Accounts*	5	5
Undermoderation*	6	11
Overmoderation*	2	3
Lockout and Control	3	3
Surveillance	1	2
False reporting	6	6

Patterns – excluded options examples

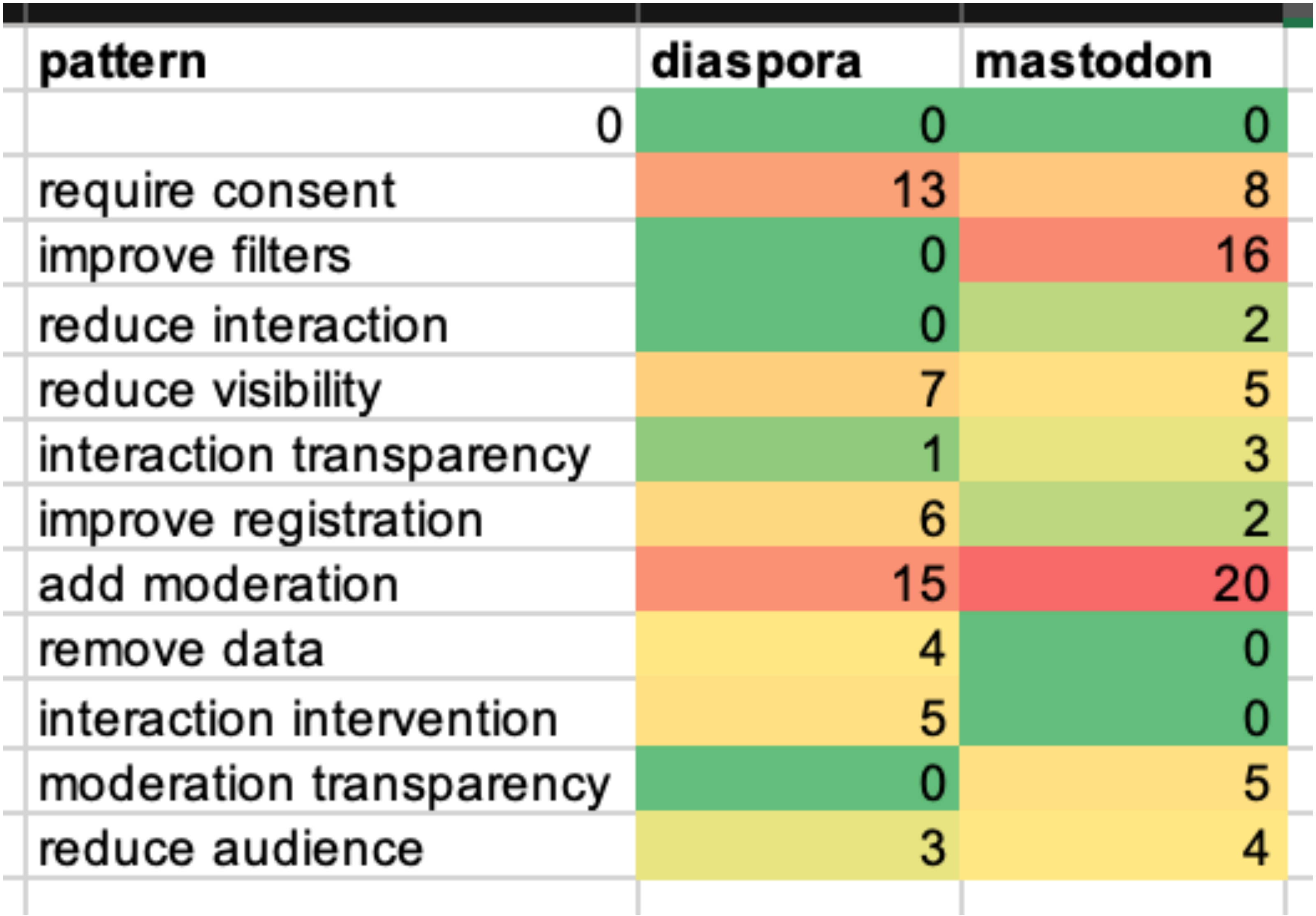
- "private messaging would be helpful"
- "I would suggest more [profile] bio space"
- "When someone #tags their comments, they should be visible on a #tag page"
- "Another solution could be if the username isn't found on the [instance], we make a search to other[s]"
- "Or even better, a choice of email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

Patterns Over Time



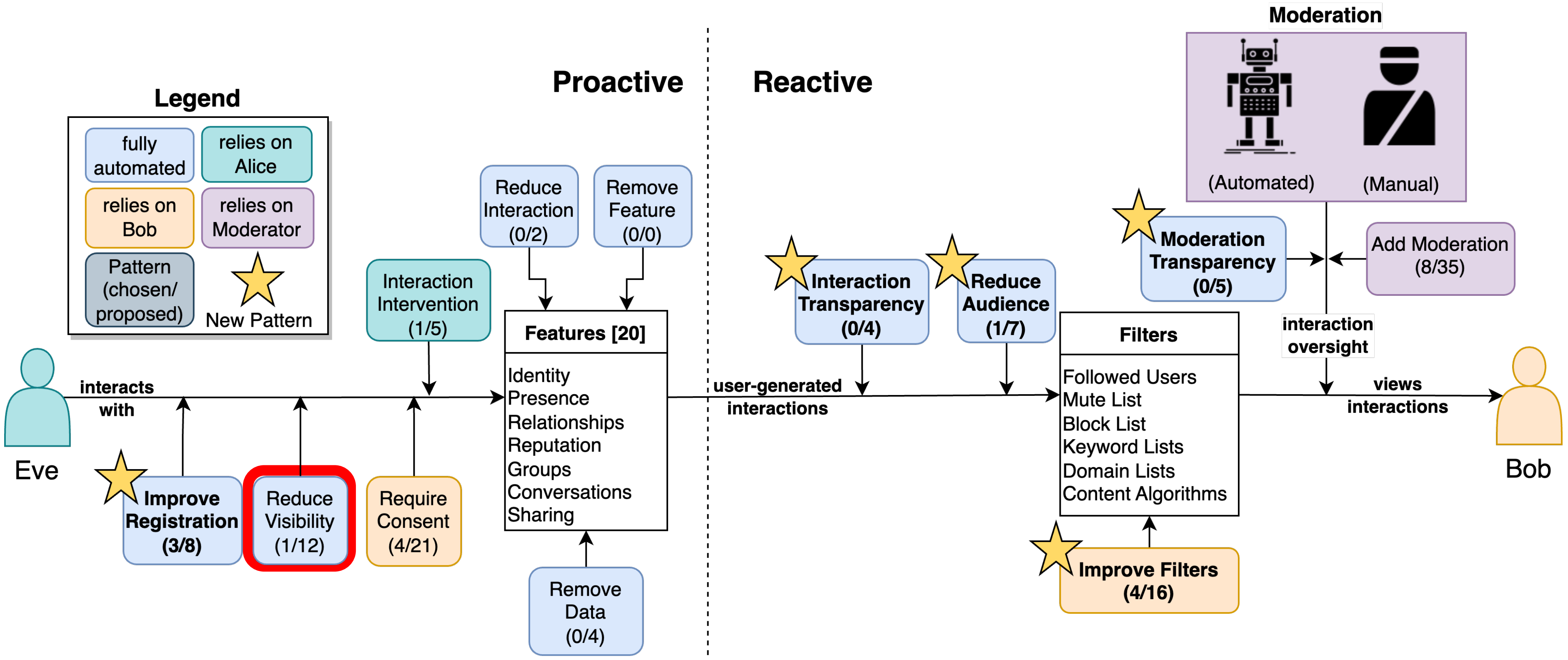
pattern	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
0	0	0	0	0	0	0	0	0	0	0	0	0	0
require consent	0	6	2	5	0	0	1	1	4	1	1	0	0
improve filters	0	0	0	0	0	0	10	4	2	0	0	0	0
reduce interaction	0	0	0	0	0	0	1	0	1	0	0	0	0
reduce visibility	0	1	0	0	3	0	1	3	3	1	0	0	0
interaction transparency	0	0	0	0	1	0	0	1	2	0	0	0	0
improve registration	0	3	0	2	0	0	0	0	1	0	1	0	1
add moderation	0	1	0	2	4	2	0	5	14	2	3	2	0
remove data	0	1	0	0	0	1	0	0	1	1	0	0	0
interaction intervention	0	0	1	0	2	1	0	1	0	0	0	0	0
moderation transparency	0	0	0	0	0	0	0	1	4	0	0	0	0
reduce audience	0	0	0	0	0	0	0	1	5	1	0	0	0

Patterns – Diaspora vs. Mastodon

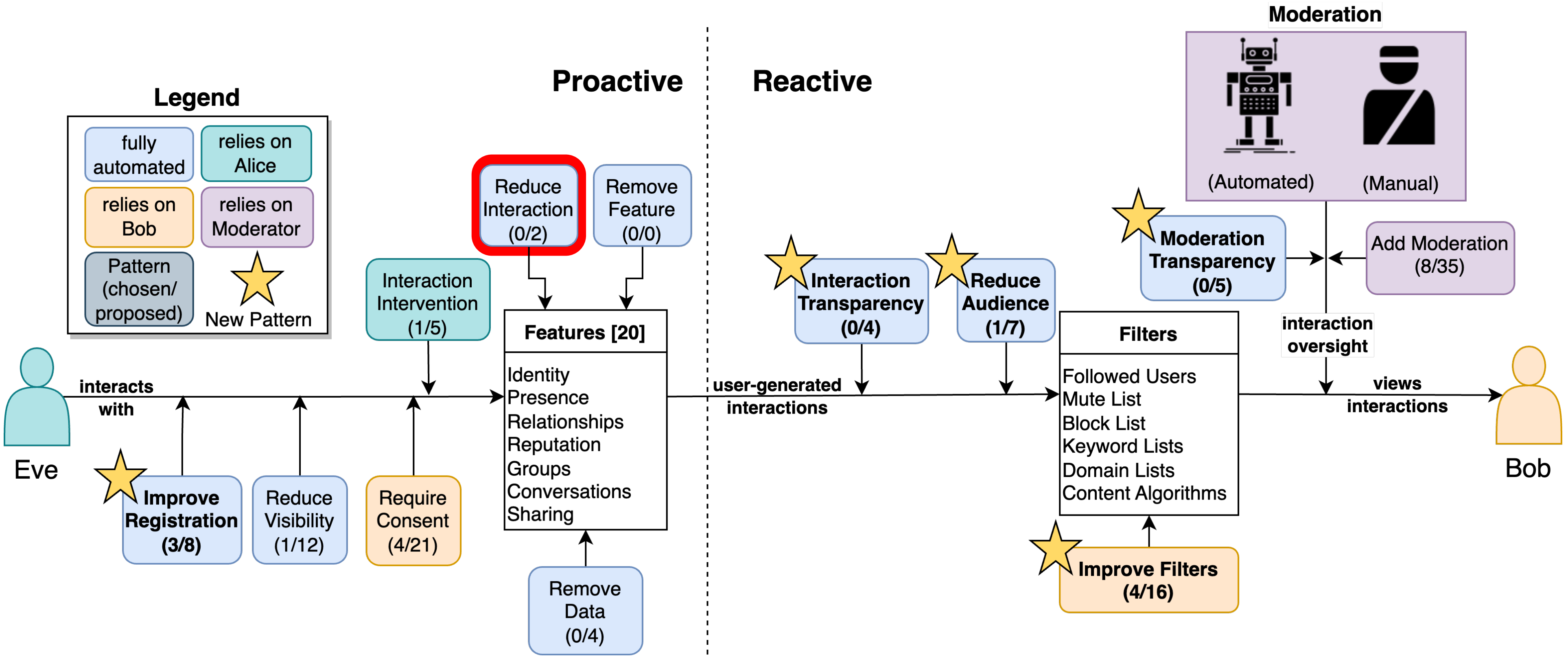


pattern	diaspora	mastodon
0	0	0
require consent	13	8
improve filters	0	16
reduce interaction	0	2
reduce visibility	7	5
interaction transparency	1	3
improve registration	6	2
add moderation	15	20
remove data	4	0
interaction intervention	5	0
moderation transparency	0	5
reduce audience	3	4

T&S Engineering Pattern Diagram



T&S Engineering Pattern Diagram



Legend

fully automated

relies on Alice

relies on Bob

relies on Moderator

Pattern (chosen/ proposed)

New Pattern

Proactive

Reactive

Moderation

(Automated)

(Manual)

Eve

interacts with

Improve Registration (3/8)

Reduce Visibility (1/12)

Require Consent (4/21)

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)

user-generated interactions

Interaction Transparency (0/4)

Reduce Audience (1/7)

Filters

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

Improve Filters (4/16)

Moderation Transparency (0/5)

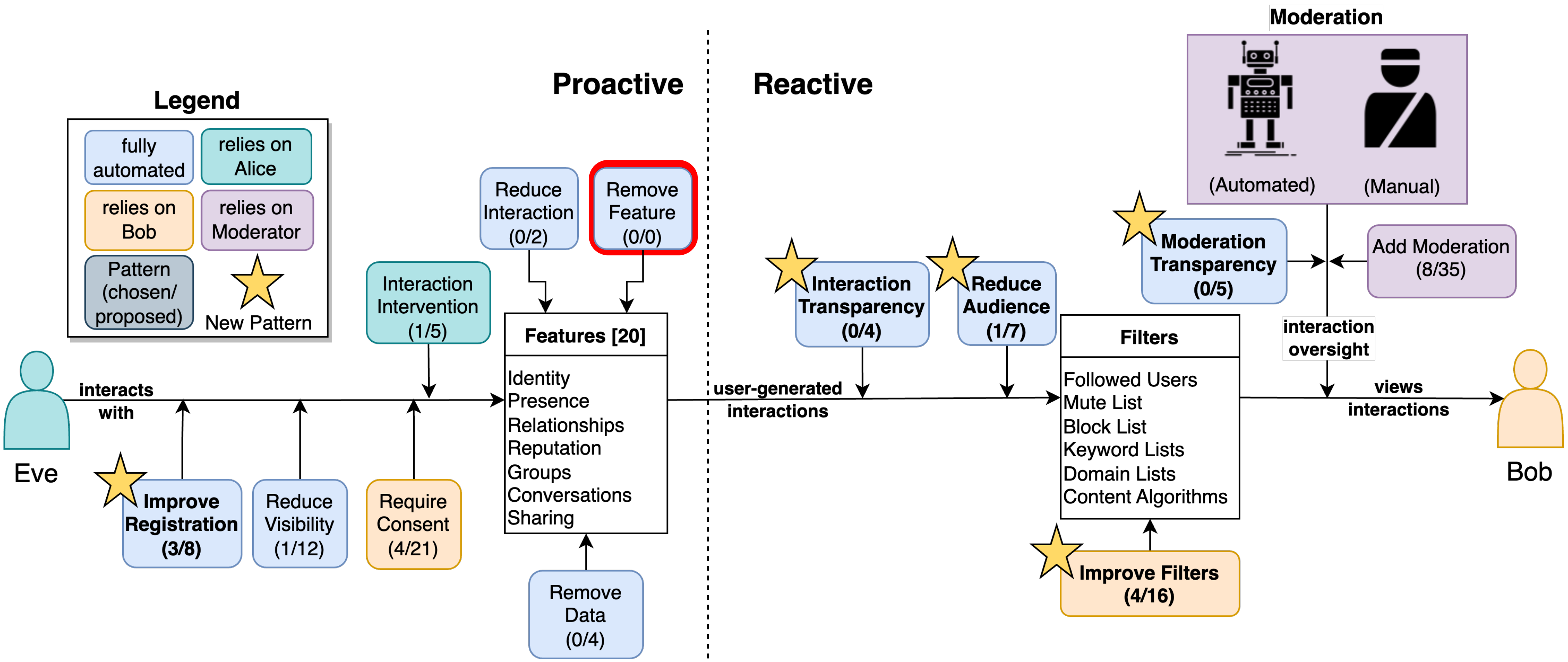
Add Moderation (8/35)

interaction oversight

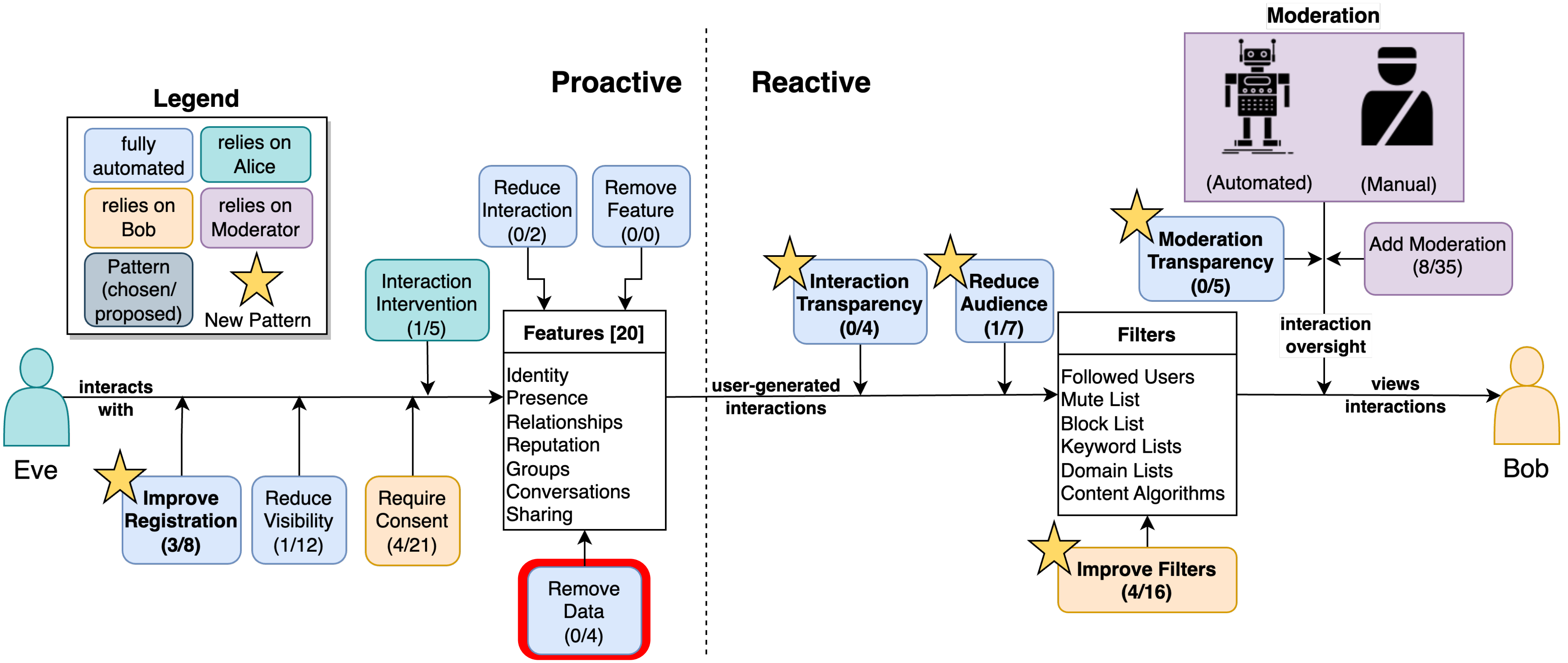
views interactions

Bob

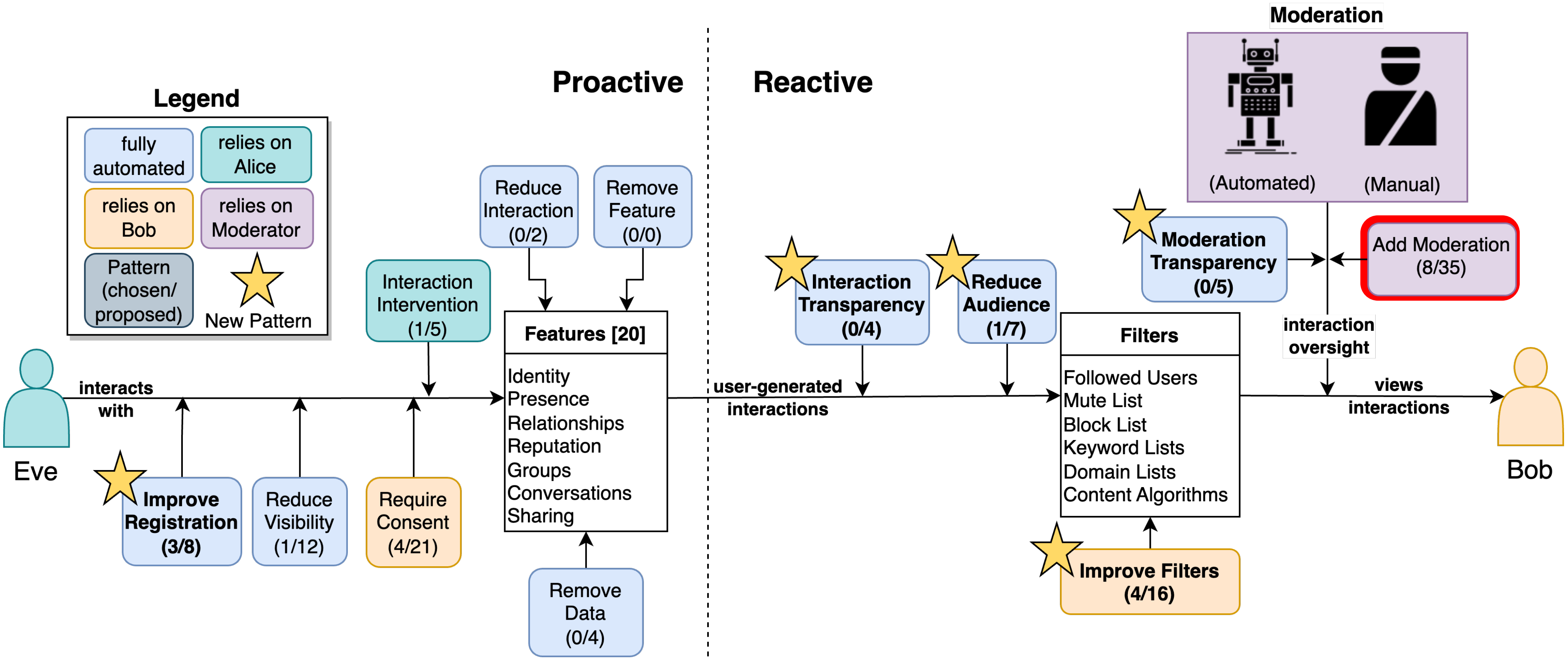
T&S Engineering Pattern Diagram



T&S Engineering Pattern Diagram



T&S Engineering Pattern Diagram



Legend

fully automated

relies on Alice

relies on Bob

relies on Moderator

Pattern (chosen/ proposed)

New Pattern

Proactive

Reactive

Moderation

(Automated)

(Manual)

Eve

interacts with

Improve Registration (3/8)

Reduce Visibility (1/12)

Require Consent (4/21)

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)

user-generated interactions

Interaction Transparency (0/4)

Reduce Audience (1/7)

Filters

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

Improve Filters (4/16)

Moderation Transparency (0/5)

Add Moderation (8/35)

interaction oversight

views interactions

Bob

Data Availability

- Artifact location: 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