Stop Reinventing in Isolation

Bringing Open Source to Trust & Safety Infrastructure

roost

Content Warning

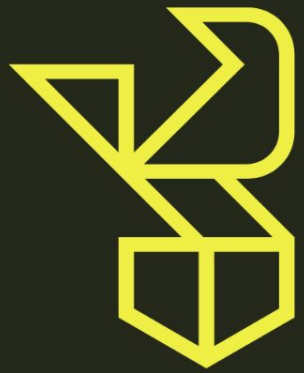
roost

Stop Reinventing in Isolation

1. Who are we?
2. T&S crash course
3. Non-profit, open source approach
4. Osprey: high performance rules engine
5. Get Involved

Who are we?

R	Robust
O	Open
O	Online
S	Safety
T	Tools



roost

Who are we?

Cassidy James Blaede
Open Source Community Manager

Anne Bertucio
Chief Operating Officer

Trust & Safety Crash Course

roost

T&S Crash Course

Imagine a world where cats are evil…

- Identify and block or deprioritize discussion of cats
- Prevent distribution of photos of cats
- Adapt your approach as cat-related harms evolve and change

More formally: DIRE Framework

Detection: identify photos and discussions of cats

Investigation: find patterns in cat-related attacks or actors

Review: assess flagged content or edge cases against cat policies

Enforcement: timeout or ban cat accounts; report to the cat authorities

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5369158

Regulatory & Reporting Requirements

Depend greatly on platform type, jurisdiction,
& local laws.

Platforms are constantly reinventing tools

- Signals lists/banks
- Detection tools (e.g. for hash matching)
- Investigation tools
- Review dashboards
- Reporting flows

Replace “cats” with other harms:

- Child sexual abuse materials (CSAM)
- Terrorist or violent extremism content (TVEC)
- Nihilistic violent extremism (NVE)
- Sadistic online exploitation(SOE)
- Scams and fraud
- Spam
- Non-consensual intimate imagery (NCII)
- Grooming
- Illegal content
- Self harm (including eating disorders)

Resources for learning more

Trust and Safety Professionals Association (TSPA)

TSPA Curriculum: tspa.org/curriculum/ts-curriculum

TSPA Library: tspa.org/explore/trust-safety-library

Non-profit & Open Source Approach

roost

Status Quo: T&S is kind of broken

- Smaller platforms and newer protocols **might not know** they need these tools
- Industry tools are **unaffordable** or **inaccessible**
- Lengthy **application & approval process** from tool owners

- Platforms **reinvent basic functionality** instead of focusing on harms specific to their community
- Generative AI and LLMs are **accelerating the production** and spread of harms
- **Openness is taboo**: “If we talk about it, bad actors will figure out how to get around our protections!”

What we believe

The missing piece is tangible tools (e.g. code)

Tools ~~can~~ *must* be open source to be accessible, modifiable, auditable, and part of a public commons

ROOST is open source!

Both licensing & development model

Hasher-Matcher-Actioner (HMA): community & office hours

ROOST Model Community (RMC): ways to use and improve open-weight models in T&S

ROOST itself: roadmap, community documentation, governance

Coop: review console we acquired and are open sourcing

Osprey: automated rules engine originating from Discord

github.com/roostorg

Osprey

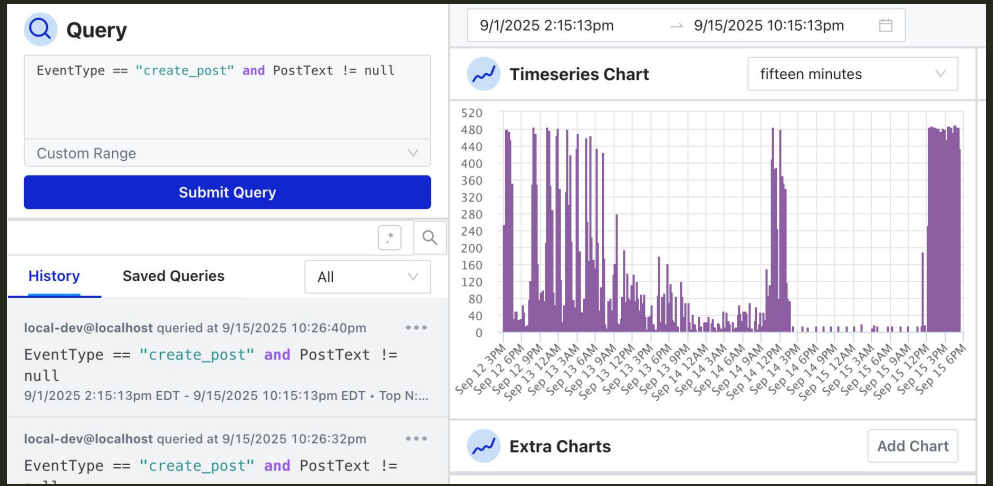
Automate the obvious,
investigate the ambiguous

roost

Osprey: Automated rules engine

Originated at **Discord**; open sourced by **ROOST**

- Built for incident response
- Self-hosted on your own infra
- Highly scalable w/built-in load balancing
- Ingests metadata eg image/video hashes, text content, IP addresses, headers, etc.



roost

When Osprey is used

Emerging incident involving cats…

Maybe word of a cat manifesto being disseminated

Accounts are linking to a cat-related site in their bio and spamming people

Accounts are joining public channels and spamming photos of cats

Determine jurisdiction for the incident to aid in reporting

Search for manifesto title or key phrases to create new labels, rules, & automations

Deep, technical analysis and investigation (IP addresses, HTTP headers, unique signatures/patterns, timestamps)

Demo!

Osprey

Query

Last Day

Submit Query

History | Saved Queries | All

local-dev@localhost queried at 1/26/2026 7:41:09pm
PostContainsSpamLink == True
Last Day • 1/25/2026 7:41:02pm EST - 1/26/2026 7:41:02pm E...

local-dev@localhost queried at 1/26/2026 7:41:06pm
PostContainsSpamLink == True
Last Day • 1/25/2026 7:41:02pm EST - 1/26/2026 7:41:02pm E...

local-dev@localhost queried at 1/26/2026 7:41:02pm
PostContainsSpamLink == True
Last Day • 1/25/2026 7:41:02pm EST - 1/26/2026 7:41:02pm E...

local-dev@localhost queried at 1/26/2026 7:40:55pm
MessageContainsSpamLink == True
Last Day • 1/25/2026 7:40:55pm EST - 1/26/2026 7:40:55pm E...

local-dev@localhost queried at 1/26/2026 7:40:41pm
SuspiciousDisplayName == True
Last Day • 1/25/2026 7:40:41pm EST - 1/26/2026 7:40:41pm E...

local-dev@localhost queried at 1/26/2026 7:39:53pm
SuspiciousDisplayName == True
Last Day • 1/25/2026 7:38:40pm EST - 1/26/2026 7:38:40pm E...

local-dev@localhost queried at 1/26/2026 7:39:46pm
SuspiciousDisplayName == True
Last Day • 1/25/2026 7:38:40pm EST - 1/26/2026 7:38:40pm E...

Timeseries Chart

hour

No data available

Extra Charts

Add Chart

Top N Results

Add Table

No data

Event Stream

Select Summary Features

No data

Osprey is production ready

Used in production by **Discord** and **Bluesky**

Being used by **Matrix.org** to bring trust and safety tooling to the open source, decentralized network

Bluesky: **over 45 million** events per day, with **over 100,000** daily enforcement actions from first set of automated rules

Discord: **orders of magnitude more** events being processed

Get Involved!

We cannot do this alone

Website & all the links:
roost.tools

Code: github.com/roostorg

Discord (we know…):
discord.gg/5Csqnw2FSQ

Try using Osprey, HMA, or RMC models and share your feedback with us!

Join office hours & working group calls

Build integrations with open platforms and protocols ❤️

Thank you!

github.com/roostorg roost.tools hello@roost.tools