Welcome

The Future of Trust & Safety:

Child Safety Investigations

Sergio Zaragoza

Child Safety Intelligence
Analyst & Consultant

Alexandra Koptyaeva

Trust and Safety
Representative at Falkor

Lior Mordechai

Customer Success &
Marketing Director at Falkor

Agenda

01 The Importance of Investigating Child Safety Issues

02 New Regulations

03 Common Challenges Across T&S Teams

04 Methodology Tips

05 Technology for the Modern Analyst

06 Q&A

The Importance of Investigating Child Safety Issues

The Importance of Investigating Child Safety Issues

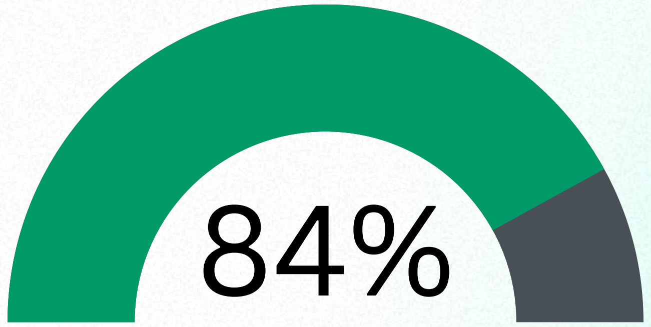
- Affects platform reputation
- Impacts user retention
  - Negative user experience
- Soon to be legally regulated globally
  - Risk assessment
  - Policy Management and Community Guidelines update
  - Compliance: know the laws, your obligations, and consequences

Child Safety is not only about **CSAM**.

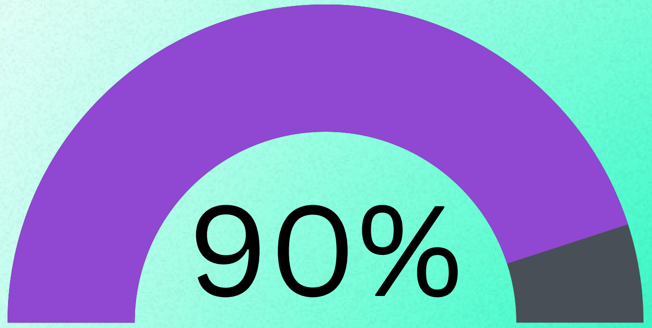
There's also **Grooming**, **Sextortion**, **Self-Generated Content**, **Deepfakes**, **Non-Consensual Indecent Imagery**, **Child Sex Trafficking**, **Child Labor Trafficking**, etc.

2022 Statistics

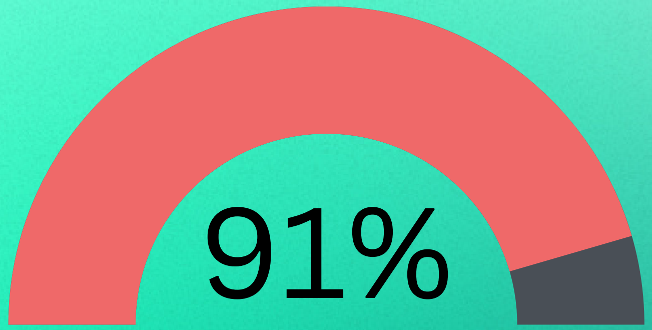
- 84% of CSAM reports were of never-before-seen material, depicting new child victims of sexual abuse.
- 9 in 10 of the victims depicted are aged 3 to 13 years old and 1% are in the infant category (0-2 years), with children as young as 4 months old victimized.
- 91% of reported CSAM involved female victims.



Never-before-seen material



Victims aged 3 to 13

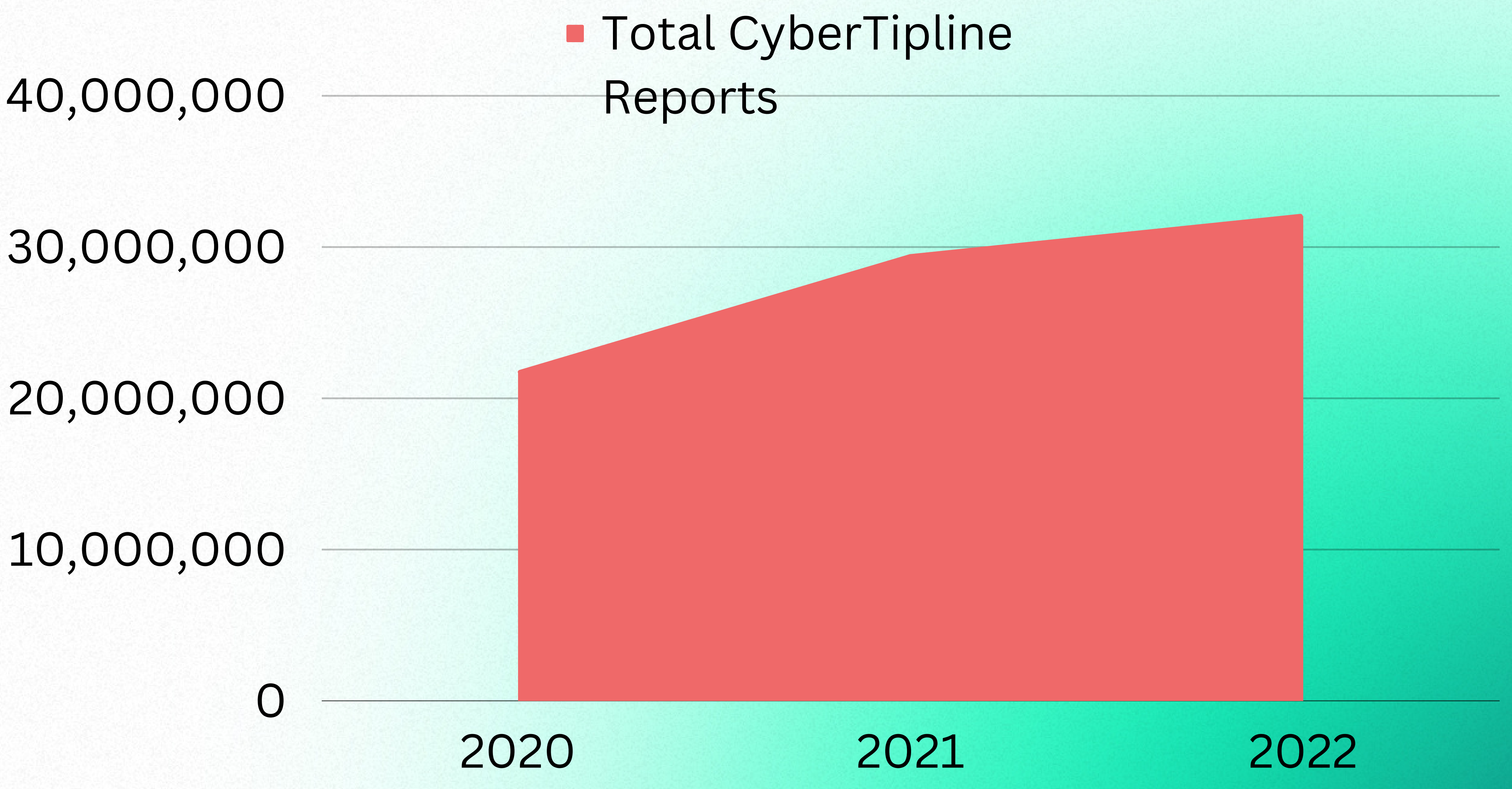


Female victims

Source: INHOPE Annual Report, 2022

2022 Statistics

Increase in CyberTipline Reports in the US:



Source: NCMEC Annual Report, 2022

CyberTipline Reports Made Available to Law Enforcement

Law Enforcement Recipients of CyberTipline Reports	2022
Internet Crimes Against Children (ICACs) Units	Actionable: 491,655 Informational: 400,715 Total: 892,370
Local Law Enforcement	Actionable: 1,462 Informational: 3 Total: 1,465
Federal Law Enforcement	Actionable: 1,356,988 Informational: 997,475 Total: 2,354,463
International Law Enforcement	Actionable: 13,995,567 Informational: 15,010,413 Total: 29,005,980

2022 Trends

Generative AI	CAP Sites	Live Streaming
• Creating and sharing (fake) CSAM imagery (e.g., AI-generated body + real victim's face). • Engaging in Grooming conversations. • Sextortion, Deepfakes. • Training LLMs with AI-generated CSAM images so they can have their own image data sets. Recent FBI warning about GenAI	• Hotlines have observed the rise of Child Abuse Pyramid Sites. ○ New type of commercial site. • Uses a particular form of “invitation” system to access CSAM content. • Users are encouraged to share their personal link to invite others to the site. • The more invitations they share, the more points they accumulate to access CSAM content.	• Individuals pay to watch live abuse of a child via a video streaming service • Incredibly difficult to detect due to real-time nature and lack of digital evidence left behind after the crime

New Regulations

More and More Regulations

1. Online Safety Bill (UK)
2. Online Safety and Media Regulation Act (Ireland)
3. Digital Services Act (Europe)
4. US state-specific regulations (e.g., A.B. 587 in California, HB 20 in Texas, S.B. 7072 in Florida)

US Regulations: AB-2273 CA Act

California Age-Appropriate Design Code Act

- Signed on **Sep 15, 2022.**
- Takes effect on **July 1, 2024.**
- New legal obligations on companies with respect to online products and services that are “likely to be accessed by children” under the age of 18.
- These platforms must comply with this Act and demonstrate that they are taking steps to protect children.

Violators may be subject to a penalty of **up to $2,500** per affected child for each negligent violation, and **up to $7,500** per affected child for each intentional violation.

US Regulations: 18 USC 2258A - NCMEC & CSAM

- Duty to Report:
    - To reduce and prevent the online sexual exploitation of children, a provider shall report as reasonably as possible after obtaining actual knowledge and take action.
    - They must inform CyberTipline.

- Failure to Report:
    - Knowingly and willfully failing to make a report — up to **$150,000** (initial knowing and willful failure);
    - Second or subsequent knowing and willful failure to make a report — up to **$300,000**.

Common Challenges Across T&S Teams

The Current State of T&S Investigations

Basic workflow:

1. Content is detected by AI & human Content Moderators
2. Content is removed according to guidelines
3. Offending Profile/Account is banned and reported to LE

But! The end problem is not solved.
Bad actors are still on the platform, and are presumably active on other platforms.

Methodology Tips

Methodology Tips - Theory

Improve the team's OSINT capabilities & general investigation skills

Keep up with current and emerging trends

Maintain an internal knowledge base of platform trends and user behavior

Attend courses that are relevant to the field

Methodology Tips - Practice

Increase your proactiveness in detecting threats (stress testing, red teaming)

Dive deep into the data (rather than focusing on one user/bad actor)

Collaborate with other teams and companies (AI, CM, specialized tools)

Increase transparency & data accessibility to relevant stakeholders

The Proactive Approach

A proactive approach is defined by **leading investigations on the platform**, so we can detect new threats and bad actors ahead of time.
We can follow **leads** that will help us to understand bad actors' **behaviour** with anticipation and act more efficiently independently of AI moderation.

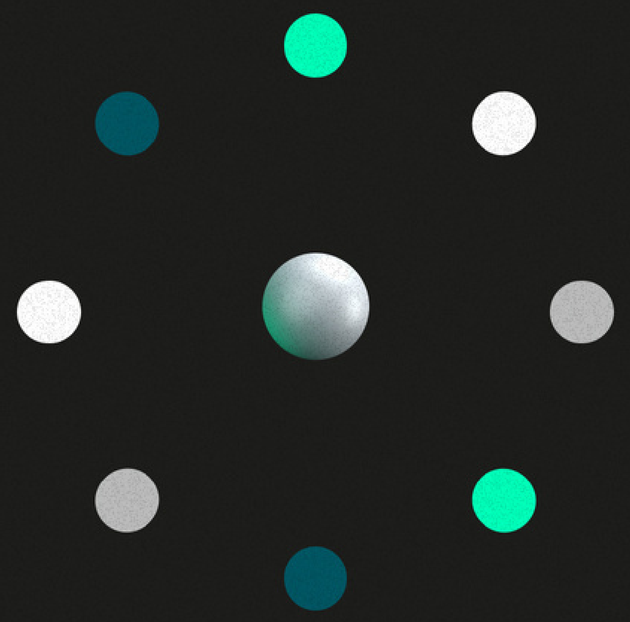
The Importance of a Proactive Approach

- Depending on the company, it can be an internal process or outsourced.
- Provides the opportunity to detect new trends and threat actors if appropriate tools and resources are available.
- By following different trails, we are able to detect future harms and behaviors before they occur.
- Platform-independent investigations can give a broader perspective on where threats may originate.
- An out-of-the-box approach to gain a deeper understanding of today's landscape.

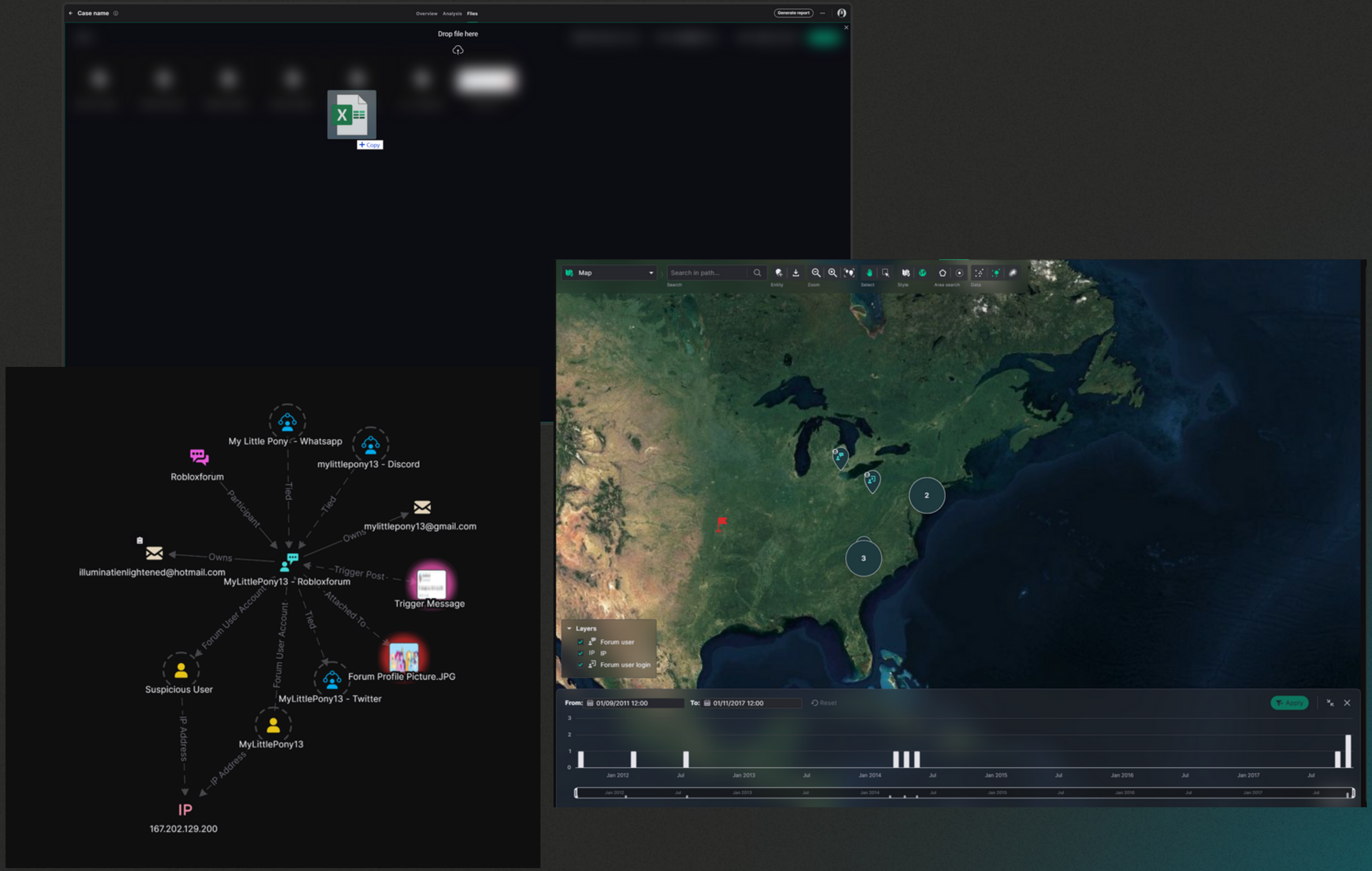
Technology for the Modern Analyst

Deeper Investigations with Falkor

- Analyze available data on problematic users
- Handle all cases in one, shared knowledgebase
- Get clearer insights into user behavior:
  - What content has been shared
  - When, how, and where
  - Possible connected accounts
- Escalating to Law Enforcement:
  - Provide more precise and concise information
  - Reduce their workload so more escalations can be addressed
  - Help speed up the investigations to help more victims



Supercharge Data Analysis



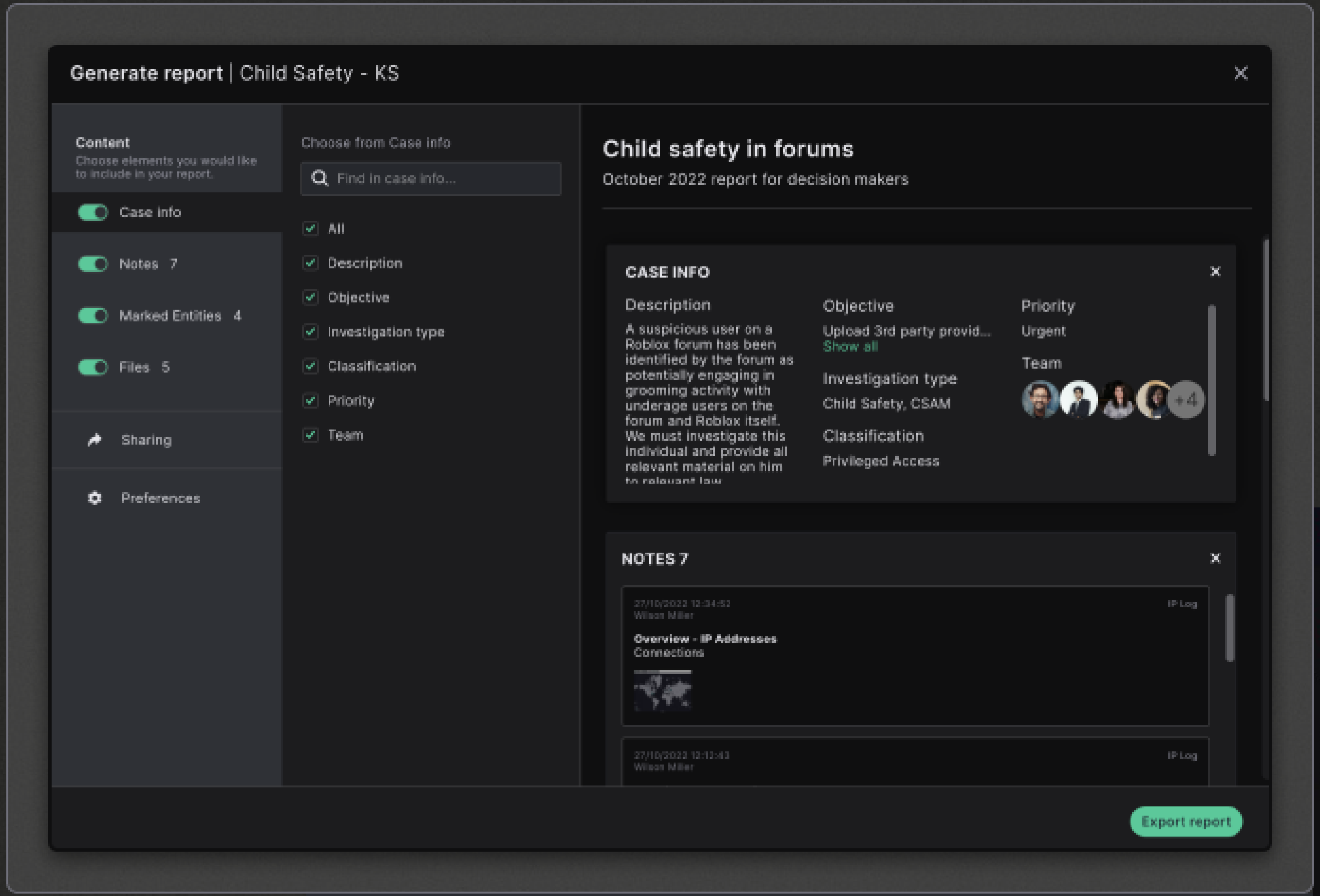
From: 01/09/2011 12:00
To: 01/11/2017 12:00

- Upload any data
- Visualize it in tables, maps, link analysis graphs, & timelines
- Filter it by time, location, or other attributes
- Analyze patterns
- Expand networks

Streamline Collaboration

- Solve cases together
- Get insights from your teammates
- Share suspicious entities
- Assign tasks and deadlines
- Automatically export reports for escalation

Streamline Collaboration



Generate report | Child Safety - KS

Content
Choose elements you would like to include in your report.

- Case info
- Notes 7
- Marked Entities 4
- Files 5

Choose from Case info

☑ All
☑ Description
☑ Objective
☑ Investigation type
☑ Classification
☑ Priority
☑ Team

Child safety in forums
October 2022 report for decision makers

CASE INFO

Description
A suspicious user on a Roblox forum has been identified by the forum as potentially engaging in grooming activity with underage users on the forum and Roblox itself. We must investigate this individual and provide all relevant material on him to relevant law

Objective
Upload 3rd party provid...
Show all

Investigation type
Child Safety, CSAM

Classification
Privileged Access

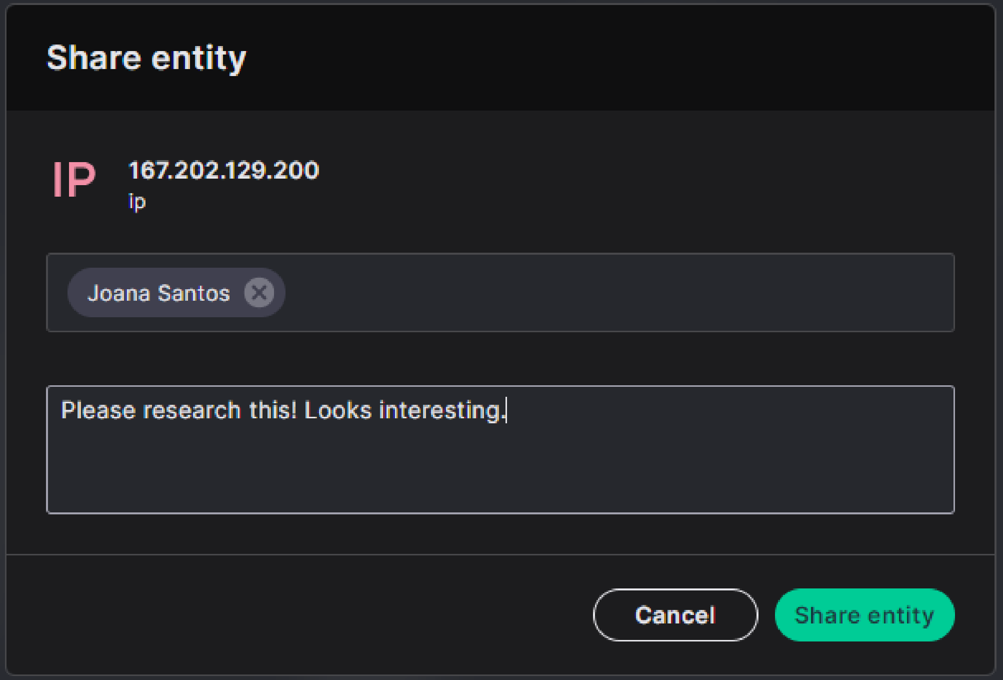
Priority
Urgent

Team
+4

NOTES 7

27/10/2022 12:34:52
Wilson Miller
IP Log
Overview - IP Addresses
Connections

27/10/2022 12:12:43
Wilson Miller
IP Log

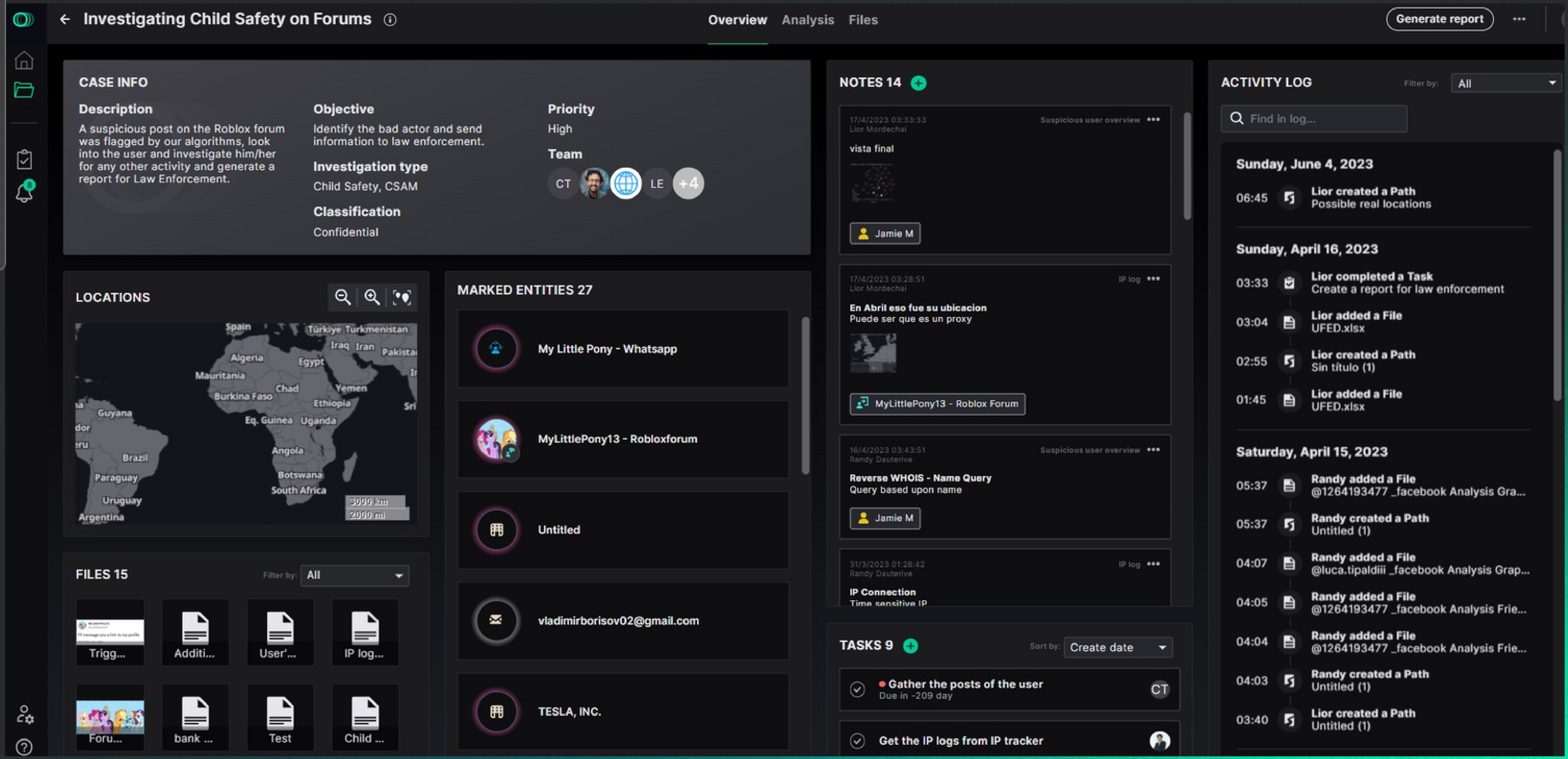


Share entity

IP 167.202.129.200
ip

Joana Santos

Please research this! Looks interesting.



Investigating Child Safety on Forums

CASE INFO

Description
A suspicious post on the Roblox forum was flagged by our algorithms, look into the user and investigate him/her for any other activity and generate a report for Law Enforcement.

Objective
Identify the bad actor and send information to law enforcement.

Investigation type
Child Safety, CSAM

Classification
Confidential

Priority
High

Team
CT LE +4

LOCATIONS

FILES 15

Filter by: All

MARKED ENTITIES 27

- My Little Pony - Whatsapp
- MyLittlePony13 - Robloxforum
- Untitled
- vladimirborisov02@gmail.com
- TESLA, INC.

NOTES 14

17/4/2023 03:33:33
Lior Mordechai
Suspicious user overview
vista final
Jamie M

17/4/2023 03:28:51
Lior Mordechai
IP log
En Abril eso fue su ubicacion
Puede ser que es un proxy
MyLittlePony13 - Roblox Forum

16/4/2023 03:43:51
Randy Dauterive
Suspicious user overview
Reverse WHOIS - Name Query
Query based upon name
Jamie M

31/3/2023 01:28:42
Randy Dauterive
IP log
IP Connection
Time sensitive IP

TASKS 9

Sort by: Create date

- Gather the posts of the user
  Due in -209 day
  CT
- Get the IP logs from IP tracker

ACTIVITY LOG

Filter by: All

Sunday, June 4, 2023

06:45 **Lior created a Path**
Possible real locations

Sunday, April 16, 2023

03:33 **Lior completed a Task**
Create a report for law enforcement

03:04 **Lior added a File**
UFED.xlsx

02:55 **Lior created a Path**
Sin título (1)

01:45 **Lior added a File**
UFED.xlsx

Saturday, April 15, 2023

05:37 **Randy added a File**
@1264193477 _facebook Analysis Gra...

05:37 **Randy created a Path**
Untitled (1)

04:07 **Randy added a File**
@luca.tipaldiii _facebook Analysis Grap...

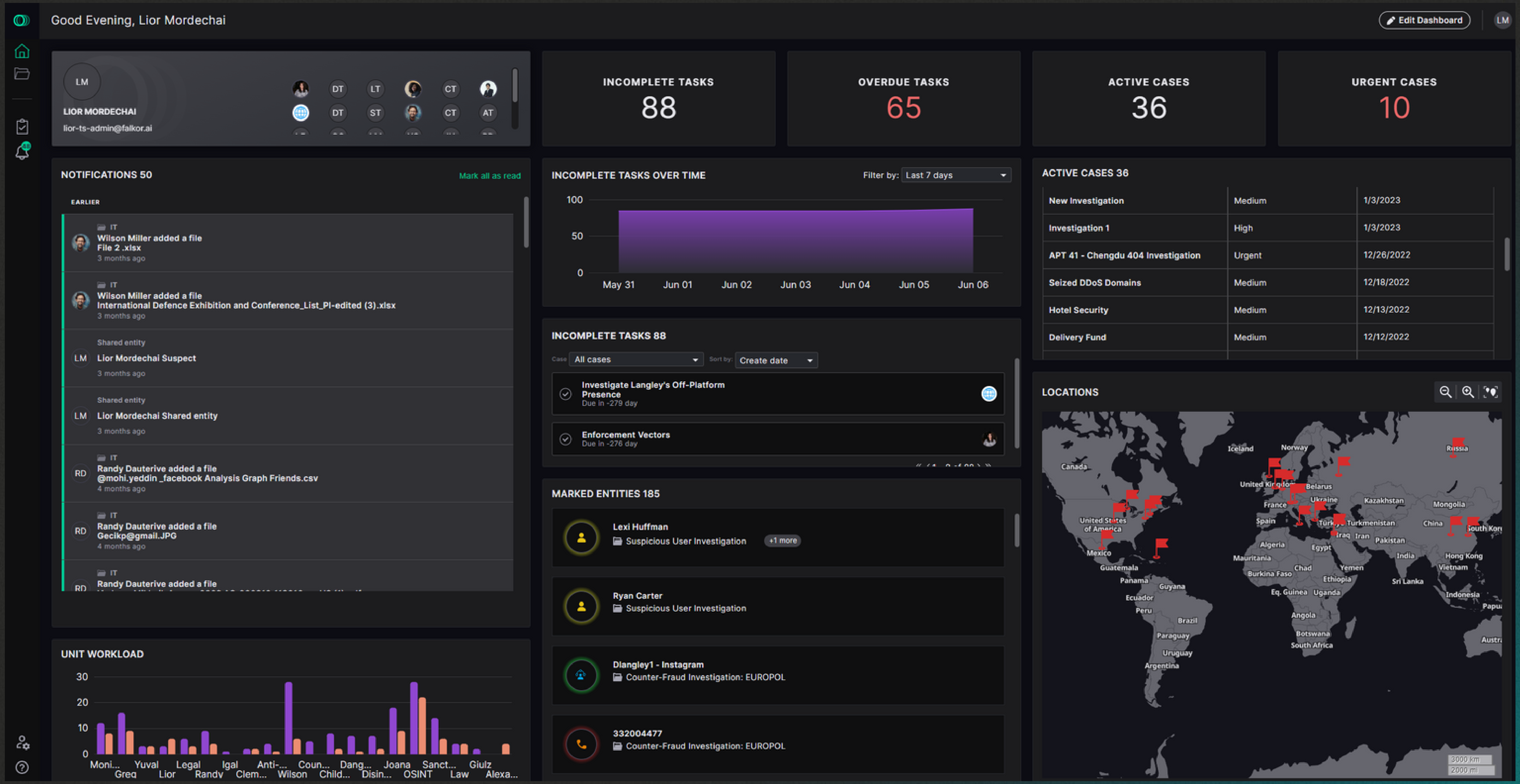
04:05 **Randy added a File**
@1264193477 _facebook Analysis Frie...

04:04 **Randy added a File**
@1264193477 _facebook Analysis Frie...

04:03 **Randy created a Path**
Untitled (1)

03:40 **Lior created a Path**
Untitled (1)

Manage Teams



Good Evening, Lior Mordechai

LM

LIOR MORDECHAI

lior-ts-admin@falkor.ai

INCOMPLETE TASKS 88

OVERDUE TASKS 65

ACTIVE CASES 36

URGENT CASES 10

NOTIFICATIONS 50

EARLIER

IT
Wilson Miller added a file
File 2 .xlsx
3 months ago

IT
Wilson Miller added a file
International Defence Exhibition and Conference_List_PI-edited (3).xlsx
3 months ago

Shared entity
LM Lior Mordechai Suspect
3 months ago

Shared entity
LM Lior Mordechai Shared entity
3 months ago

IT
RD Randy Dauterive added a file
@mohi.yeddin _facebook Analysis Graph Friends.csv
4 months ago

IT
RD Randy Dauterive added a file
Gecikp@gmail.JPG
4 months ago

IT
RD Randy Dauterive added a file

INCOMPLETE TASKS OVER TIME

Filter by: Last 7 days

INCOMPLETE TASKS 88

Case: All cases
Sort by: Create date

Investigate Langley's Off-Platform Presence
Due in -279 day

Enforcement Vectors
Due in -276 day

MARKED ENTITIES 185

Lexi Huffman
Suspicious User Investigation +1 more

Ryan Carter
Suspicious User Investigation

Dlangley1 - Instagram
Counter-Fraud Investigation: EUROPOL

332004477
Counter-Fraud Investigation: EUROPOL

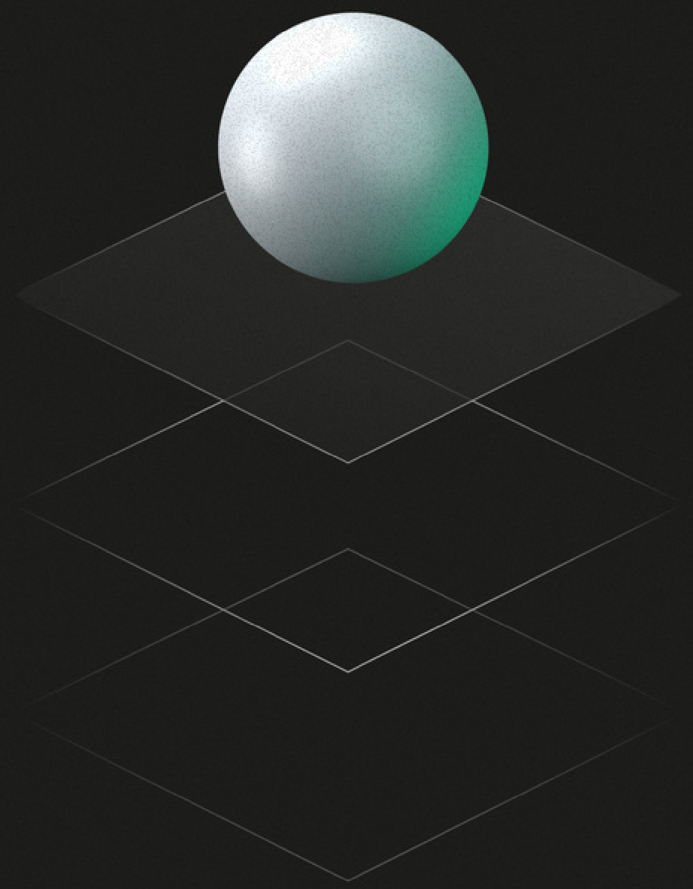
ACTIVE CASES 36

New Investigation	Medium	1/3/2023
Investigation 1	High	1/3/2023
APT 41 - Chengdu 404 Investigation	Urgent	12/26/2022
Seized DDoS Domains	Medium	12/18/2022
Hotel Security	Medium	12/13/2022
Delivery Fund	Medium	12/12/2022

LOCATIONS

UNIT WORKLOAD

- Get a bird's eye view of your team's work
- Ensure efficiency and employee wellness
- Manage permissions and roles
- Audit activity and improve processes



As a Trust & Safety professional, you do not need to investigate *everything* on the platform in-depth.

But using your data correctly can help you identify the root cause of issues, eliminate them for good, and prevent future problems.

Resources

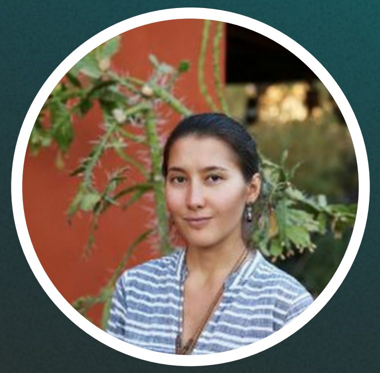
- INHOPE Annual Report, 2022
- National Center for Missing and Exploited Children (NCMEC) Transparency Report, 2022
- Sociobits, 'Dark Side of AI Art: Deepfake, Child Pornography and Sextortion', March 2023
- The Guardian, 'AI tools could be used by predators to ‘automate child grooming’, May 2023

Visit falkor.ai

Thank you!

Sergio Zaragoza

smartinezzaragoza@gmail.com



Alexandra Koptyaeva

alexandra@falkor.ai

Lior Mordechai

lior@falkor.ai