FARO

Putting AI to work on the data imbalance behind online harm

SHIELD - APS, Italy

Meheret Takele Mandefro

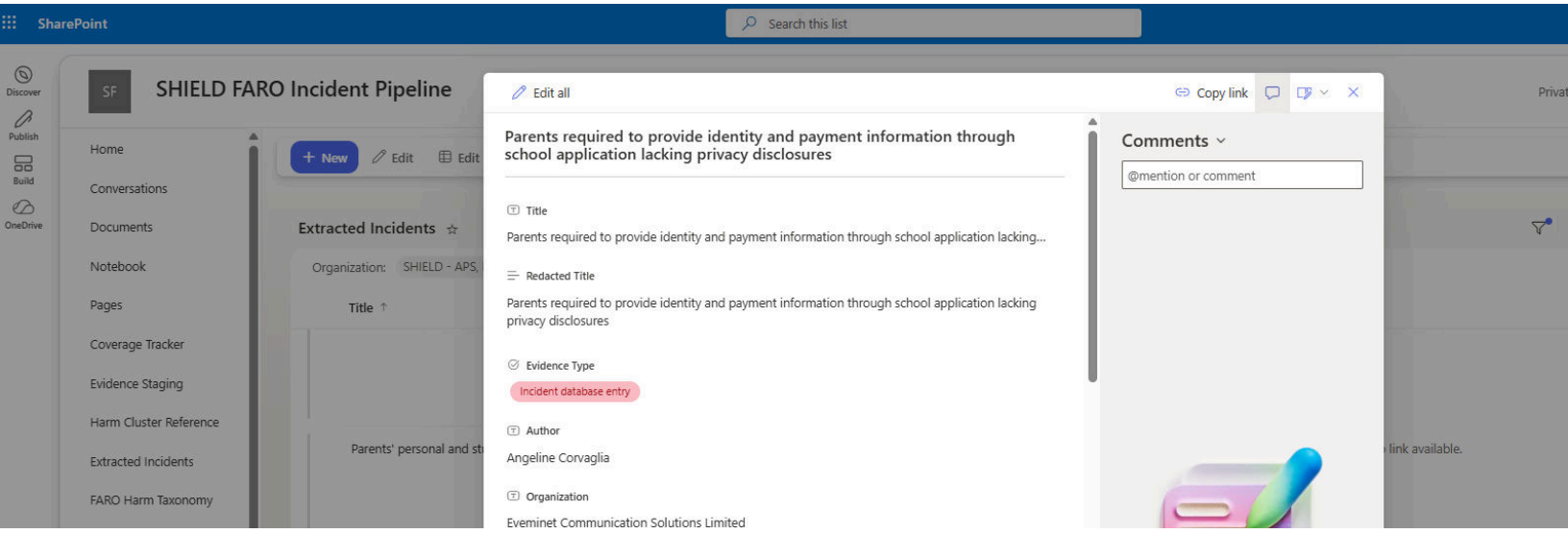
Digital Transformation Analyst, NetHope · October 2026

ABOUT NETHOPE CONNECTED

NetHope Connected is a global community and digital program designed by, for and with nonprofits, to help small to mid-sized organizations strengthen digital capacity at their own pace. It does this through learning resources, peer connection and practical guidance for adopting AI and technology responsibly.

This case study focuses on NetHope Connected's AI grant work, supporting Connectors to explore responsible, purpose-driven AI.

FARO (Frontline Aggregated Record of Online Harms) is a proof-of-concept tool built by SHIELD - APS to address a gap in how online harms are handled. The people closest to the harm rarely hold enough consolidated evidence to be heard by the decision makers who could act on it. FARO uses a chain of AI agents to gather harm reports from many sources and turn them into clear trends. It presents the findings in two versions, one written for decision makers and one for the community that supplied the information. A person reviews each agent's work before the next agent begins, and the design puts anonymity, consent and community ownership of data first.



An extracted incident record in the FARO pipeline, with its redacted title, evidence type and country.

ABOUT THE ORGANIZATION

SHIELD - APS is a global network of grassroots organizations tackling online harms across many countries. It connects grassroots organizations, practitioners and young people already working on the front lines of online safety, then helps that work reach the people and institutions positioned to act on it. The network has nearly 100 members across 24 countries and is built through direct relationships, with leadership held by the people closest to each issue.

Angeline Corvaglia
Executive Director

shieldthefuture.com
SHIELD - APS

The problem: only one side comes to the table with data

01

Almost everyone agrees that online harm is a problem, including technology companies, governments, civil society and users. Governments and regulators decide how to respond based on evidence, and technology companies arrive with plenty of it. They know who uses their products and what happens on them, and they can bring that data to any discussion.

The organizations working directly with the people being harmed arrive with much less. Their information is gathered by hand from their own communities. It is firsthand, and it records what company data does not, including the context an incident happened in and what it did to the person. It is also scattered and unstructured, and turning it into evidence that regulators will act on takes knowledge, time and money a small organization does not have. As a result, the organizations closest to the harm have the least influence on how governments and regulators respond to it.

> *"A tech company comes to a government with tons of information. A community representative comes with the little they were able to collect. There is no comparison between the two, leaving decision-makers with a far richer picture of one perspective than the other."*
>
> Angeline Corvaglia, Executive Director, SHIELD

Why AI was the right tool

AI was not chosen to make judgements about harm. It was chosen to do the work these organizations cannot afford to do by hand, which is turning large amounts of scattered information into evidence.

01

Reach across sources

The AI agent draws on public forums, news reports, official records and published research, alongside direct accounts reported to SHIELD and its partners. Bringing sources of different kinds and quality together lets a single trend be built from all of them.

02

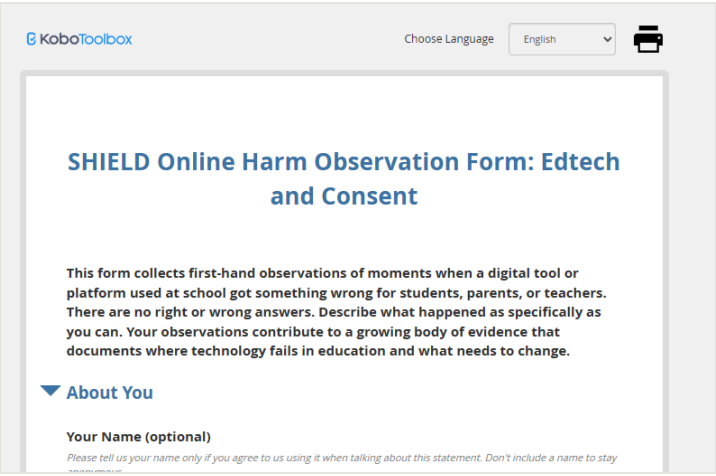
Consolidate what is scattered

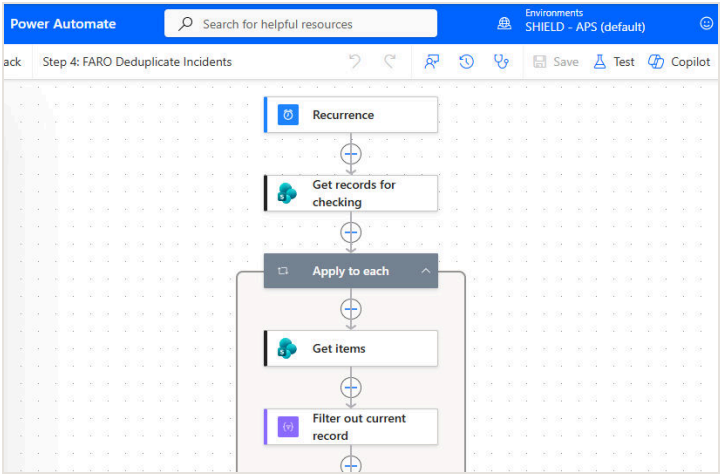
People describe what happened in conversations and interviews, not in case files. A single interview can contain several incidents mixed together. AI can read large volumes of unstructured input and pull out the individual incidents inside it.

03

Generate draft reports

The AI agent generates draft versions of each report. One is written for decision makers in government, in regulation and at the technology companies. The other is written for the community that supplied the information. A person completes each draft before it goes out.





Left: the SHIELD Online Harm Observation Form in KoboToolbox, where contributors type or record what happened. Right: one of the agent steps in the automated pipeline.

Developing FARO

02

The NetHope Connected AI grant funded a six-month project to build and test FARO as a proof of concept. Working to that timeframe, the team began by defining what the system had to deliver at the end, then worked backwards to what was needed to get there. Three outputs were fixed from the start. The first is a searchable platform where anyone can look up harms by trend or keyword. The second is reports for the bodies able to act on them, and the third is reports that go back to the communities that supplied the information.

The third output answers a problem in how harms reporting usually works. Harms reporting is often extractive, whether or not anyone intends it to be. Information is gathered, it is escalated and the people who provided it receive nothing they can use where they live. FARO makes the return report a required part of the design.

> *"People report the harms, it goes into something that gets reported up, and they never get something back that they can use locally."*
>
> Angeline Corvaglia, Executive Director, SHIELD

Who shaped the design

01

Partner organizations in the network

The team took the idea to trusted partners: Internet Safe Kids Africa in Nigeria, Eveminet Communications Solutions Limited in Kenya, Digital Wellness Africa in Zimbabwe and Social & Media Matters in India, along with RiseIQ in the UK and Internet Safety Labs in the United States. It asked whether this was the problem they faced. The partners confirmed it and raised two practical requirements. The tool had to work across languages, and it had to work both online and offline.

02

Technical and methodological advisors

The team consulted Internet Safety Labs, a research organization that specializes in online harms and in the systems used to classify them. Internet Safety Labs shared its incident classification system with SHIELD and advised on how to sort and consolidate the incidents FARO collects, so FARO's categories follow established practice.

03

Listening to those directly affected

For its first focus area, education technology, the team consulted people with direct experience of the harm. They included parents whose children had faced serious problems with education technology and youth online safety advocates from multiple countries, who could speak to both their own experiences and what they hear from the communities they support.

04

Existing research on harms reporting

When the first collection method drew little response, published research on harms reporting explained why and pointed to the alternative the team adopted.

How FARO is built

FARO runs as a chain of seven AI agents, each given a single job. The team chose seven separate agents instead of one because, in testing, an agent narrowed to one task produced more consistent results than one asked to handle several steps at once. Splitting the work this way also creates a point between each stage where the output can be checked before the next agent runs.

Information enters through KoboToolbox, a form tool built for field data collection. It works offline, supports several languages and accepts typed answers or a recorded voice note, which covers the two requirements partners raised. Longer accounts come from one-hour interviews with community representatives. Collected data is stored in the Microsoft suite, and the agents run on Google's Gemini models, accessed through Google's API.

Inside the FARO pipeline

02

The seven agents are grouped into four main stages.

01	02	03	04
Source search	**Incident extraction**	**Anonymization**	**Analysis and reports**
AI searches online by country and harm keyword and flags sources that may describe individual incidents. A person approves or rejects each source.	AI reads each approved source and pulls out every individual incident that passes the qualification test, flagging sources it cannot access. A person checks that each incident is real.	A dedicated agent replaces names with role, age band and gender, and generalizes dates and locations for minors. A person checks the redacted text and anything the agent flagged as uncertain.	AI flags possible duplicates, classifies each incident against the harm taxonomy, then drafts both versions of the report. A person reviews each step and approves each report.

No model was trained on contributor data. The agents work from written instructions, known as prompts, and from the incident classification system Internet Safety Labs shared.

Deciding what counts as acceptable evidence was a separate question. A court judgment, a regulatory decision and a public post do not carry the same weight, and the team had to decide which of them to accept. SHIELD followed an approach used by organizations such as Amnesty International and Human Rights Watch, which draw on all three and combine them to establish a pattern rather than to prove a single case.

Choosing a provider

The team had planned to build FARO on Microsoft's Copilot and switched to Google's Gemini during development, for two reasons.

Cost per call

Each report passes through seven agents, so the pipeline sends a large number of requests to the AI model. At that volume Copilot was too expensive, and Gemini was the least costly option the team found.

Access to the subject matter

General-purpose AI assistants restrict searches about the most severe harms, and those limits blocked some of the material FARO needed to find. In the team's testing, Gemini's restrictions were narrower than Copilot's.

Responsible design in practice

ANONYMITY BY DESIGN

The system is designed to show patterns across a region, so it never needs to hold the detail that would identify a specific school, community or person. Removing that detail is built into the pipeline as its own step.

CONSENT AND NON-EXTRACTION

The team does not collect posts from social media, even though they would be an easy source, because posting in public is not consent to this use. Interview participants are paid for their time, and every community that contributes receives a report it can use.

THE DATA BELONGS TO THE REPORTER

People who contribute keep ownership of what they share. SHIELD has permission to use it, consolidate it and publish reports from it on the platform, but never to sell it. In the longer term, the team hopes contributors could benefit from their data once the platform operates at scale.

HUMAN IN THE LOOP THROUGHOUT

A person checks each agent's output before the next agent runs. They confirm that an extracted incident is real, that a cited source exists and that a drafted interpretation holds. The team automates the listing of incidents and leaves their interpretation to people.

Putting FARO into practice

03

How it works today

FARO currently runs campaign by campaign rather than continuously. SHIELD or a partner organization has a question that would benefit from wider data collection and analysis than the organization itself has the capacity for, and FARO is used to collect and consolidate what that community knows. Contributors submit through the form in their own language, by typing or by voice, and community representatives are interviewed directly. FARO's agents then process the material through the four stages described above, with a person checking each one. The partner receives a short report with charts, patterns and trends. The reports are designed to be used in these ways:

FOR DECISION MAKERS

- Showing the real-world impact of new rules and regulations.
- Comparing different responses to similar harms, to assess which guidelines and knowledge would keep a response from worsening the harm itself.

FOR COMMUNITIES

- Sharing what has happened elsewhere, so people know what to look out for.
- Practical guidance on turning harm patterns and individual cases into what to watch out for and share locally to help keep people safe.

Who uses it

For now, the users are people working in online safety, responsible technology and responsible AI who already understand the problem in their context and need a way to evidence it. The scale of any campaign depends on the reach of the partner, not the tool. One example is Eveminet Communications Solutions Limited in Kenya, which works with 50 youth online safety trainers. Each trainer reaches around 100 young people, so a campaign run through that network could gather close to 5,000 contributions. No campaign has yet run at that scale.

Early response

Feedback from partners has been positive, though the team is clear about its limits. Partners have seen the approach described and examples of what a report will contain, but not a finished report. The reports are running later than planned, because the collection method changed partway through, as described under Challenges in rollout. Until partners hold a completed report, the response is encouraging but not yet conclusive.

CHALLENGES IN ROLLOUT

The data collection method that had to change

The original plan was to gather reports directly from individuals through a form. The slow community response made it clear that the chosen method would not scale, and among those who did respond, many were unclear about what counted as a harm worth reporting. The pattern held in every country. The team turned to published research on harms reporting to understand why, and found that people mostly report when they can see a benefit to themselves.

That created a circular problem. People would contribute once they could see useful reports, and the tool needed contributions to produce those reports. To break the circle, SHIELD interviewed community representatives who had already heard about these harms directly. Voice notes were added to the form at the same stage, so people who prefer to speak can report that way.

Working within general-purpose AI tools

Building the pipeline on commercial, general-purpose tools brought three constraints the team had not anticipated.

Safety filters. Even on Gemini, the agent whose job is to search published harm reports often returns nothing on the most severe forms of abuse, because the safety filters on these tools block that kind of search. The restriction is reasonable in its own context, but still an obstacle to legitimate research.

Invented links. Sometimes the agent returns a source link that looks plausible but leads nowhere, because it reconstructs web addresses rather than copying them. Changing the prompts reduced how often this happens, and checking every source catches the rest.

A two-minute timeout. The off-the-shelf automation tool that connects the agents has a built-in limit: it stops waiting for a response from the AI model after two minutes. Complex searches take longer, so some steps are run by hand. Working around a limit the team does not control needs paid specialist work, and it is the main reason the proof of concept cannot scale as built.

What the team learned

04

> *"I would have involved a designer at the beginning. It is a different mindset."*
>
> Angeline Corvaglia, Executive Director, SHIELD

What went well

Testing the assumption before building. Going to partner organizations first confirmed the problem was real and surfaced the language and offline requirements early, while they could still shape the design.

Setting the ethical boundaries before collection started. Anonymization, the decision not to collect posts from social media and the commitment never to sell contributed data were fixed at the design stage.

Giving each agent one job. Breaking the pipeline into single-purpose steps made each one testable in isolation, and made it clear where a person needed to check the output.

What they would do differently

The team offers two reflections for others considering similar work. The first is to bring someone experienced in designing automated systems into the project from the start. That person could have offered other options at points where the team made choices that later took time to work around. The two-minute timeout that now limits the pipeline is one example.

The second is to review existing research before choosing a data collection method. The limitations of collecting reports directly from individuals were already documented in published research, but the team found that research only after the approach had not worked. Knowing earlier might not have changed the decision to try it, but the team could have run both methods in parallel, with interviews ready as a backup if direct reporting did not work.

What comes next

REBUILDING TO SCALE

- Work with an experienced systems designer on a version that can scale, building on the proof of concept, removing the two-minute timeout and reducing the cost per API call.
- Test other AI services whose safety filters allow the legitimate harms research this work needs.
- Design for continuous collection, which the proof of concept supported but switched off because of the timeout and the specialist review it needs.
- Make the human check at each agent step a paid, specialist role with mental health support, given the distressing nature of the material.

COLLABORATION AND REACH

- Continue offering the tool campaign by campaign wherever a partner in the network has a question and no capacity to answer it.
- Draw on existing harms databases instead of duplicating them. One operator who records named incidents daily has agreed to let FARO include his data, with attribution.
- Work with organizations that document harms at large scale as part of a broader mandate, so that FARO's community-level evidence adds to what they already publish.

MONITORING AND SUSTAINABILITY

- Use the first reports, two high-level reports and one for each partner, to test whether evidence gathered across several countries gets a different response from decision-makers than evidence from a single community, and to assess how local reports can be designed to provide genuine value to participating communities and organizations.
- Explore paid access for organizations that can use the evidence, including researchers and companies that need to understand how their products affect people. Any income would sustain the tool and flow back to the people who contributed the data, since it belongs to them.

Learn more about NetHope Connected

NetHope Connected is a global community and digital program designed by, for and with nonprofits, to help small to mid-sized organizations strengthen digital capacity at their own pace.

The NetHope Connected AI grant program is one of the programs in the community, where we support grantees like SHIELD - APS to design, test and run AI responsibly, and publish the results as case studies like this one.

GET IN TOUCH

connected@nethope.org

JOIN THE COMMUNITY

Join NetHope Connected

SCAN TO JOIN

With thanks to Angeline Corvaglia. Developed with support from the NetHope Connected AI grant program. October 2026.