Using AI to detect, classify, and track disinformation

AI.DEFENSE.1 Stephen Campbell and Julian Neylan

EU CYBERNET SUMMER SCHOOL 2025

Cyber Crisis Management: Navigating Disinformation and Cyber Attacks in the AI Era

EU CyberNet

Federal Foreign Office

Funded by the European Union

Module Outline

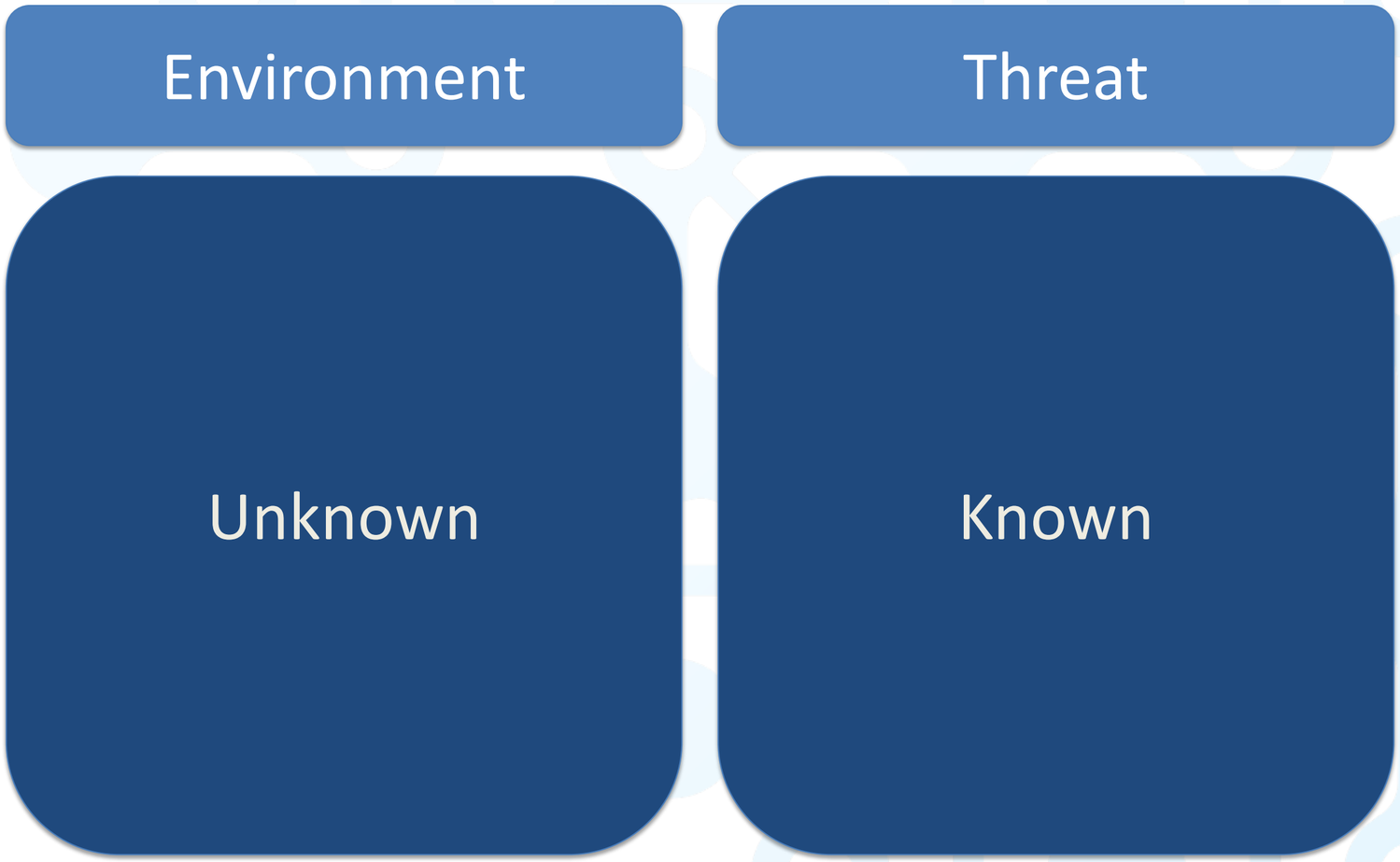
- Approaches to Detecting Disinformation
- Example AI Detection Algorithms
- Assessing Coordinated Inauthentic Behavior
- Example Investigative Tools
- Some Investigative Tool Tips

EU CyberNet
Federal Foreign Office
Funded by the European Union

Rumsfeld’s Spectrum of Uncertainty

EU CyberNet
Federal Foreign Office
Funded by the European Union

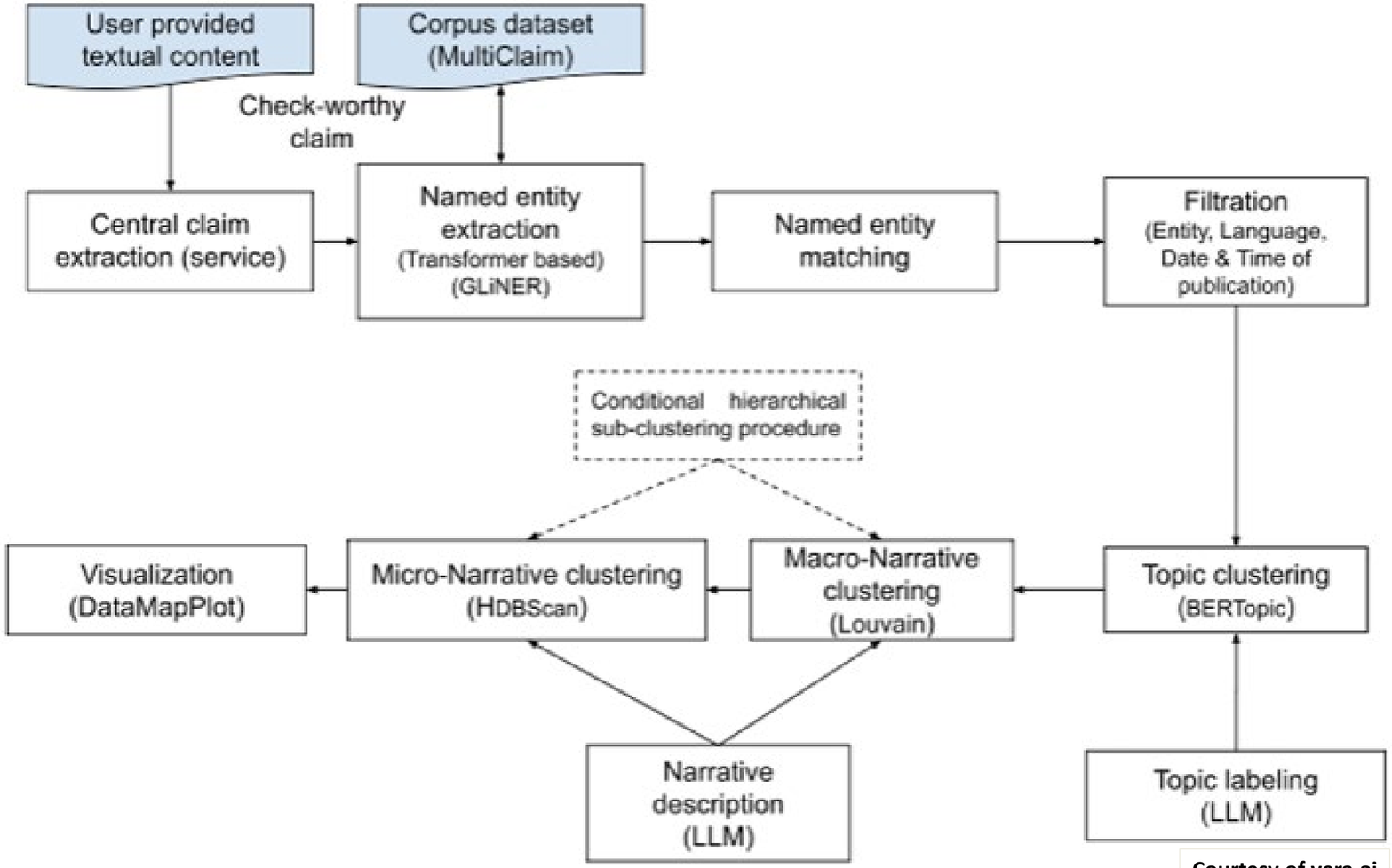
The Threat Epistemology



The Four Types of Threat Detection

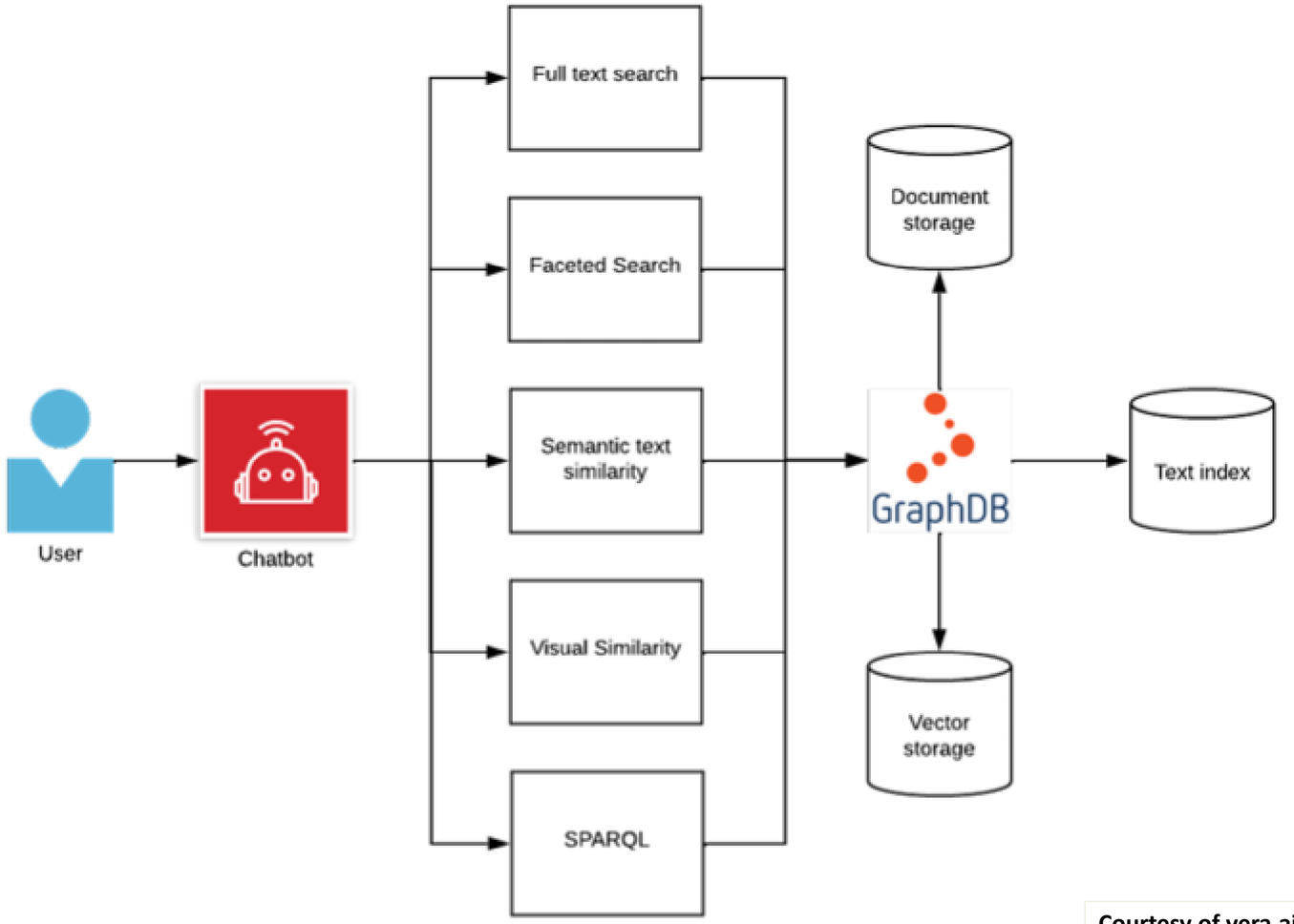
	Environment	Threat
Behaviors	Interaction-Based	Behavioral Rules
Artefacts	Similarity-Based	Indicator Rules
	Anomaly-based	Signature-based

Example Algorithms: Kempelen Institute Narrative Pipeline



Courtesy of vera.ai

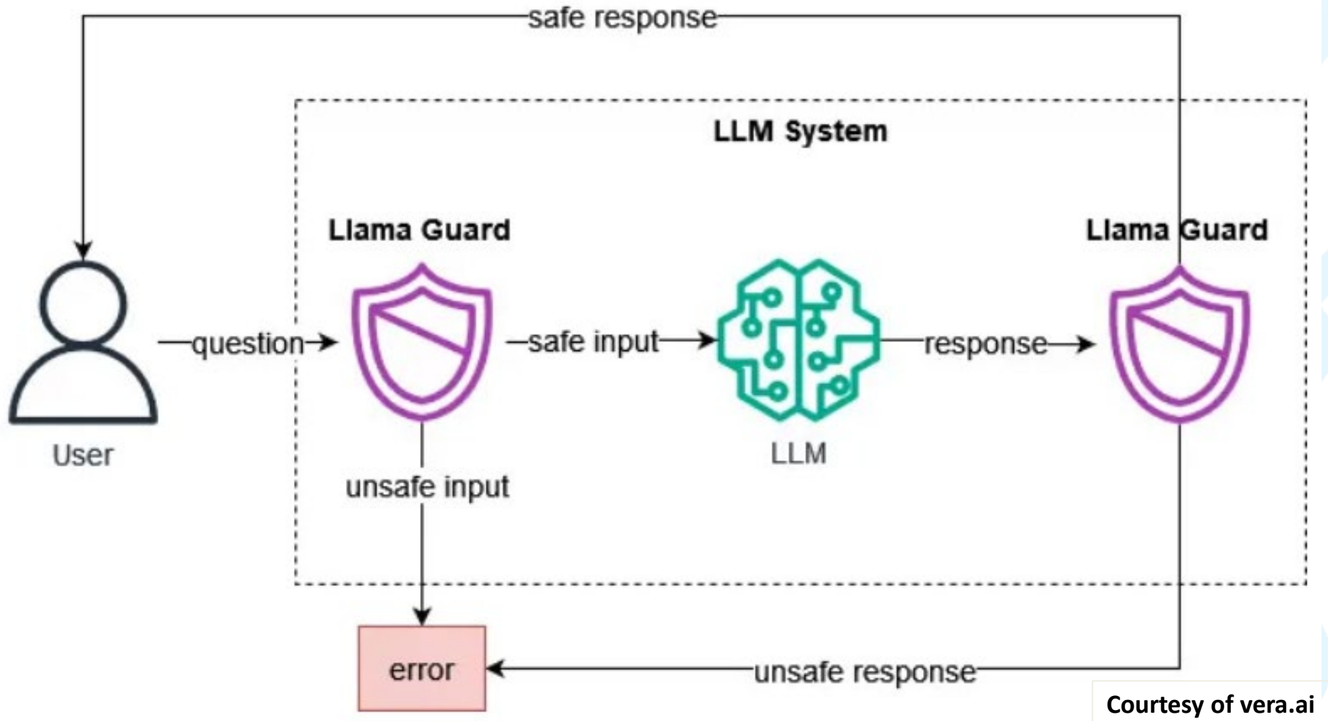
Example Algorithms: Database of Known Fakes



Courtesy of vera.ai

Approaches to Safeguarding LLMs

- Training Data Curation
- Guardrails
- LLM Firewalls
  - Meta Llama Guard
  - IBM Granite Guardian
  - CloudFlare Firewall for AI
  - Akamai Firewall for AI
- Machine Unlearning ?!



Courtesy of vera.ai

Detecting CIBs

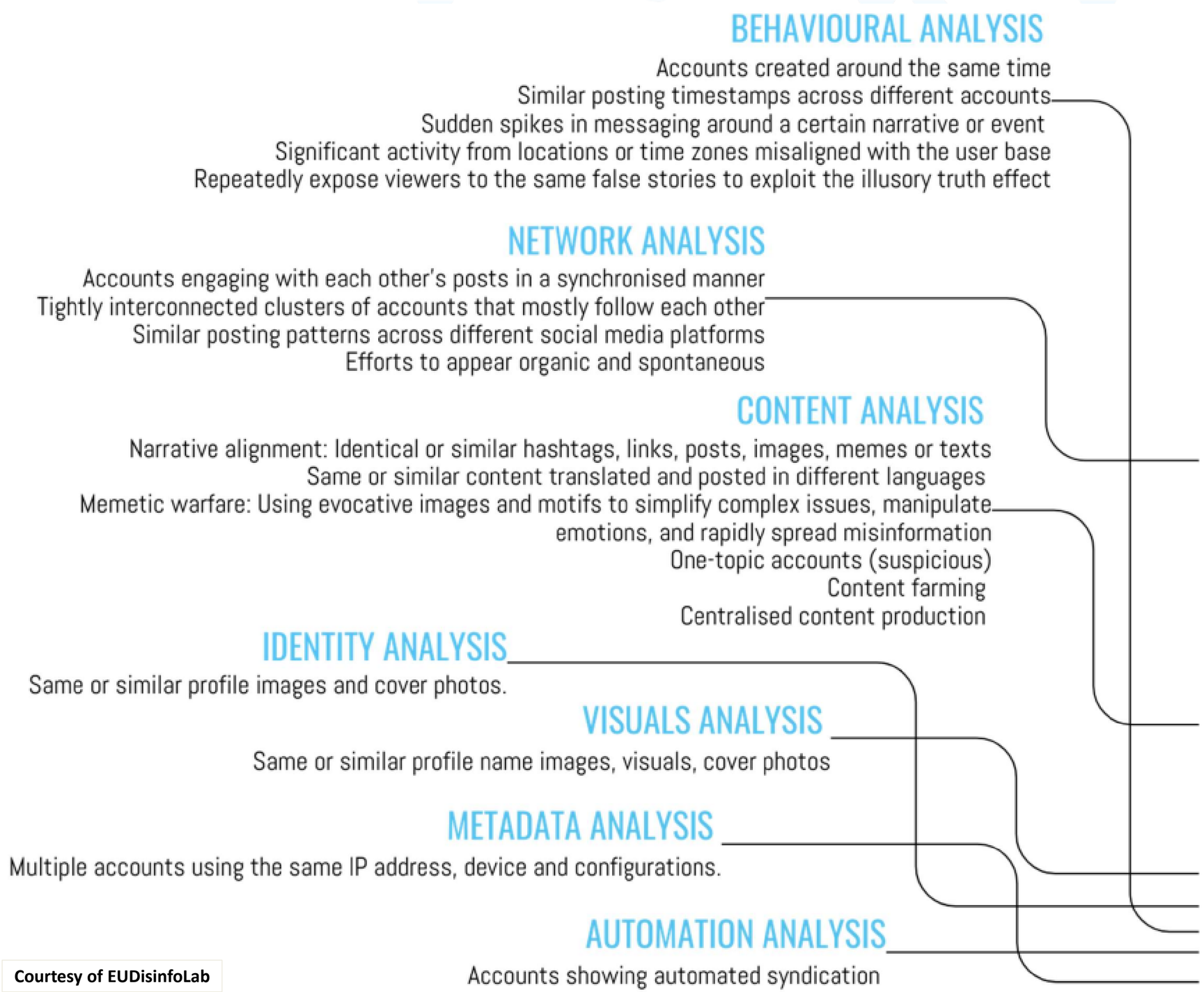
- Repost Network
  - The simplest type
- Coordinated Repost Network
  - Synchronized amplification
- Coordinated Posting Networks
  - Copypasta in action
- Coordinated Hashtag Networks
  - Hashtag hijacking
- Coordinated Hashtag Sequence Networks
  - Obfuscating Coordination
- Coordinated Mention Networks
  - Amplification and targeting through mentions
- Coordinated Link Sharing
  - Amplification through synchronized URL sharing
- Coordinated Image Sharing
  - Spamouflage
- Coordinated Account Handles
  - Handle similarity, timestamps
- Coordinated Websites
  - Interconnected domains
- Syndicated Activity
  - Temporal posting patterns

Courtesy of Jerry Gao

Example: vera.ai CIB Assessment of Operation Overload

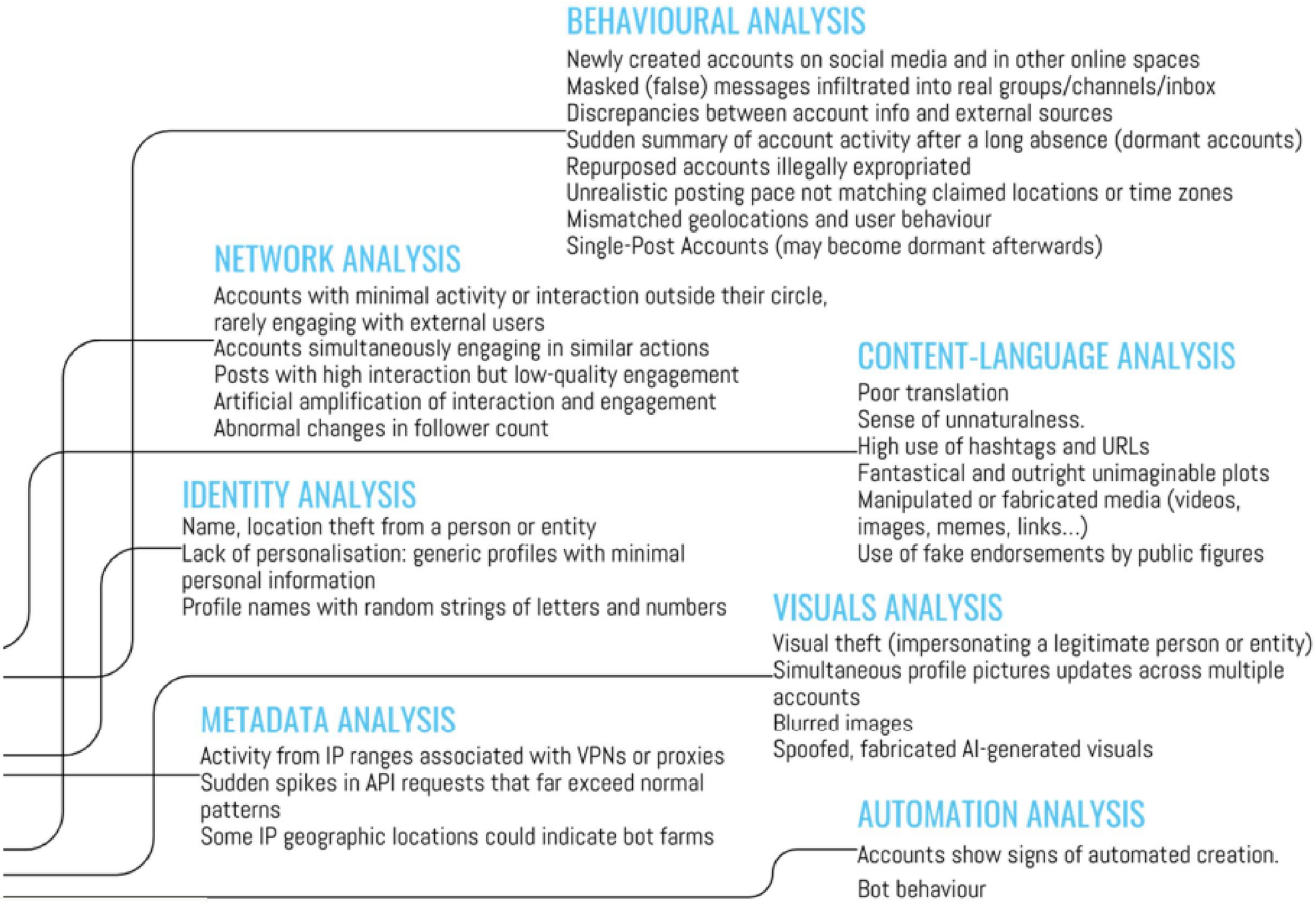
Courtesy of vera.ai

Coordination Assessment



Courtesy of EUDisinfoLab

Authenticity Assessment



Courtesy of EUDisinfoLab

Source Assessment

NETWORK ANALYSIS

Identify primary spreaders and amplifiers, often those with extensive connections or interactions
Accounts engaging first are often linked to campaign origin
Cross-platform activity can multiply leads, aiding in source tracking

CONTENT ANALYSIS

Parsing content to identify involved parties within context

IDENTITY ANALYSIS

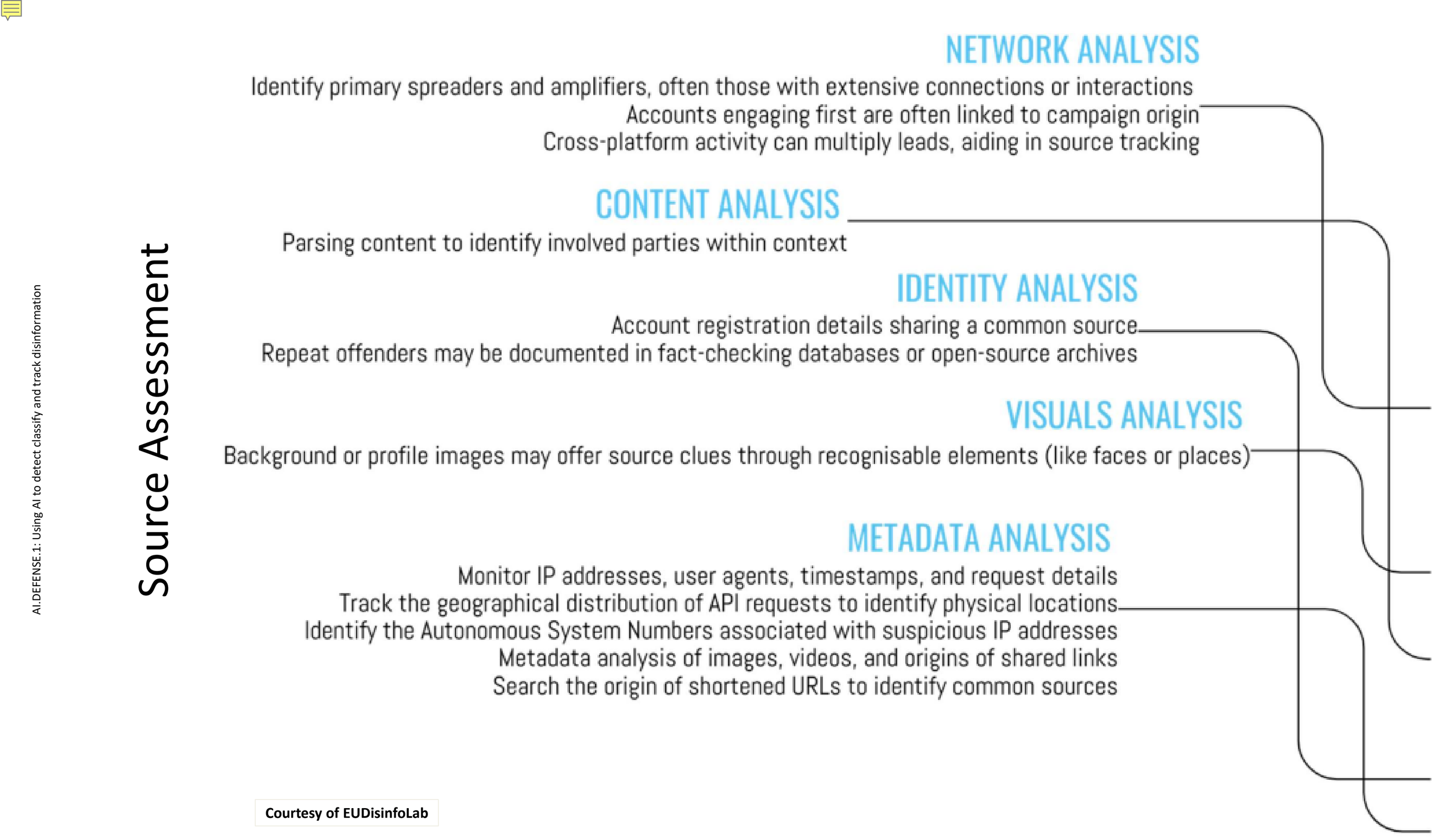
Account registration details sharing a common source
Repeat offenders may be documented in fact-checking databases or open-source archives

VISUALS ANALYSIS

Background or profile images may offer source clues through recognisable elements (like faces or places)

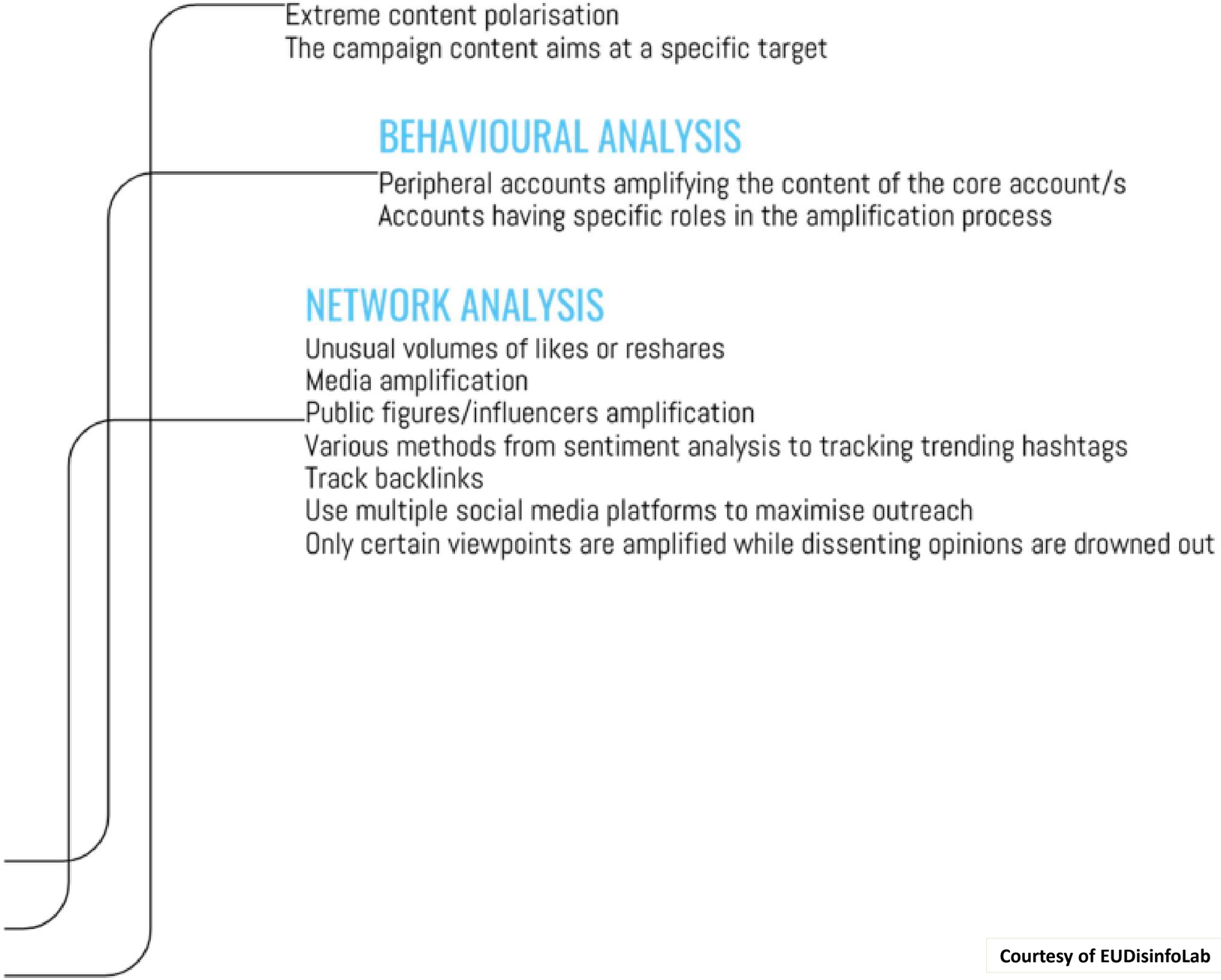
METADATA ANALYSIS

Monitor IP addresses, user agents, timestamps, and request details
Track the geographical distribution of API requests to identify physical locations
Identify the Autonomous System Numbers associated with suspicious IP addresses
Metadata analysis of images, videos, and origins of shared links
Search the origin of shortened URLs to identify common sources



Courtesy of EUDisinfoLab

Impact Assessment



Courtesy of EUDisinfoLab

Without AI you use Manual Tracking

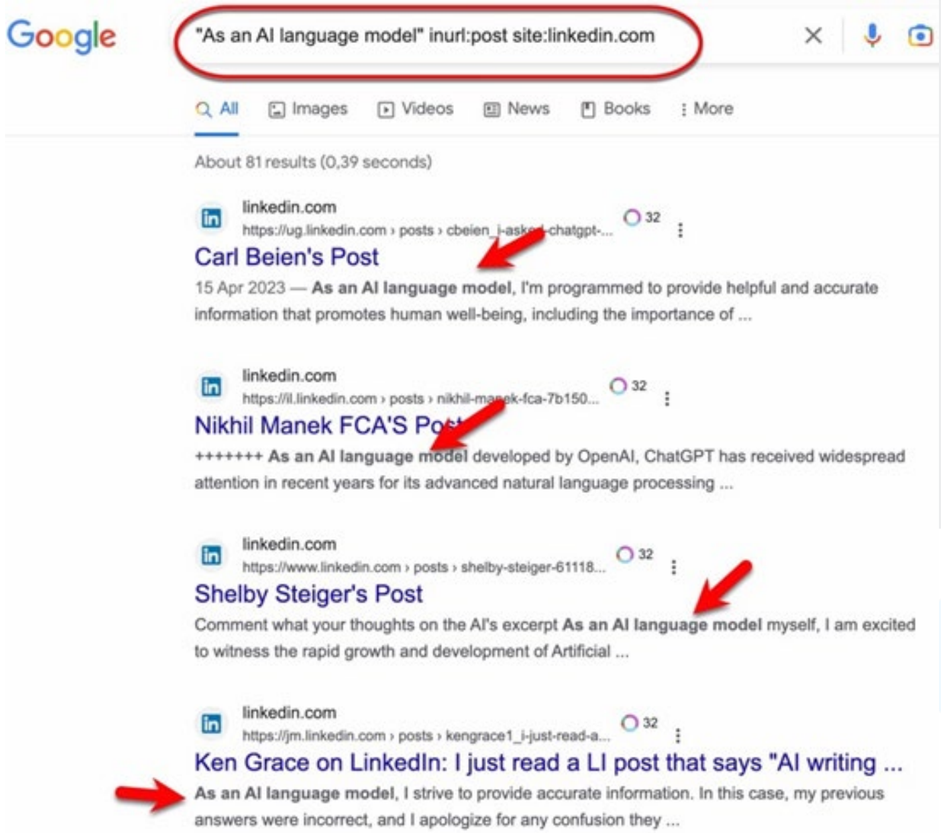
Google Searches (or Google Dorking)

Looking for Keywords and setting up the parameters for the search to give the best results

Manually looking at all of the accounts with certain content or features of interest

However there are also a broad range of tools that can be used

Example: Google Search for LLM-generated Content



- "as an AI language model"
- "not a recognized word"
- "cannot provide a phrase"
- "with the given words"
- "violates OpenAI's content policy."
- "I'm sorry, I cannot generate"
- "The message you submitted"
- "An error occurred. If this issue persists please contact us through our help center at"
- "Something went wrong, please try reloading the conversation."
- "I'm sorry, but I cannot fulfill this request"
- "Thank you for starting it" comes from "This is such an important conversation that needs to happen more often. Thank you for starting it."

Courtesy of Nico Dekens

Bellingcat toolkit

Bellingcat's Online Investigation Toolkit

Home

New Tools

CATEGORIES

- Maps & Satellites
- Geolocation
- Image/Video
- Social Media
- People
- Websites
- Companies & Finance
- Conflict

CATEGORIES

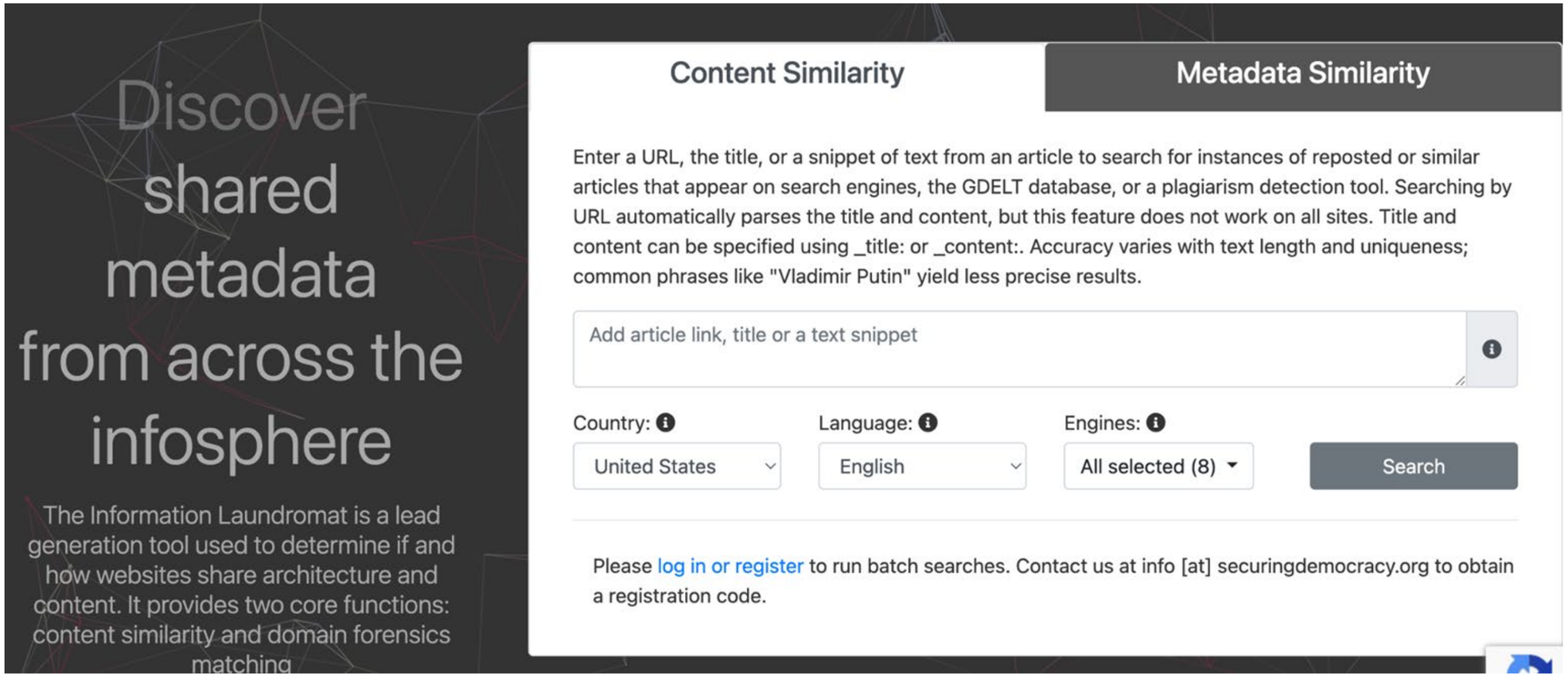
Copy

Social Media

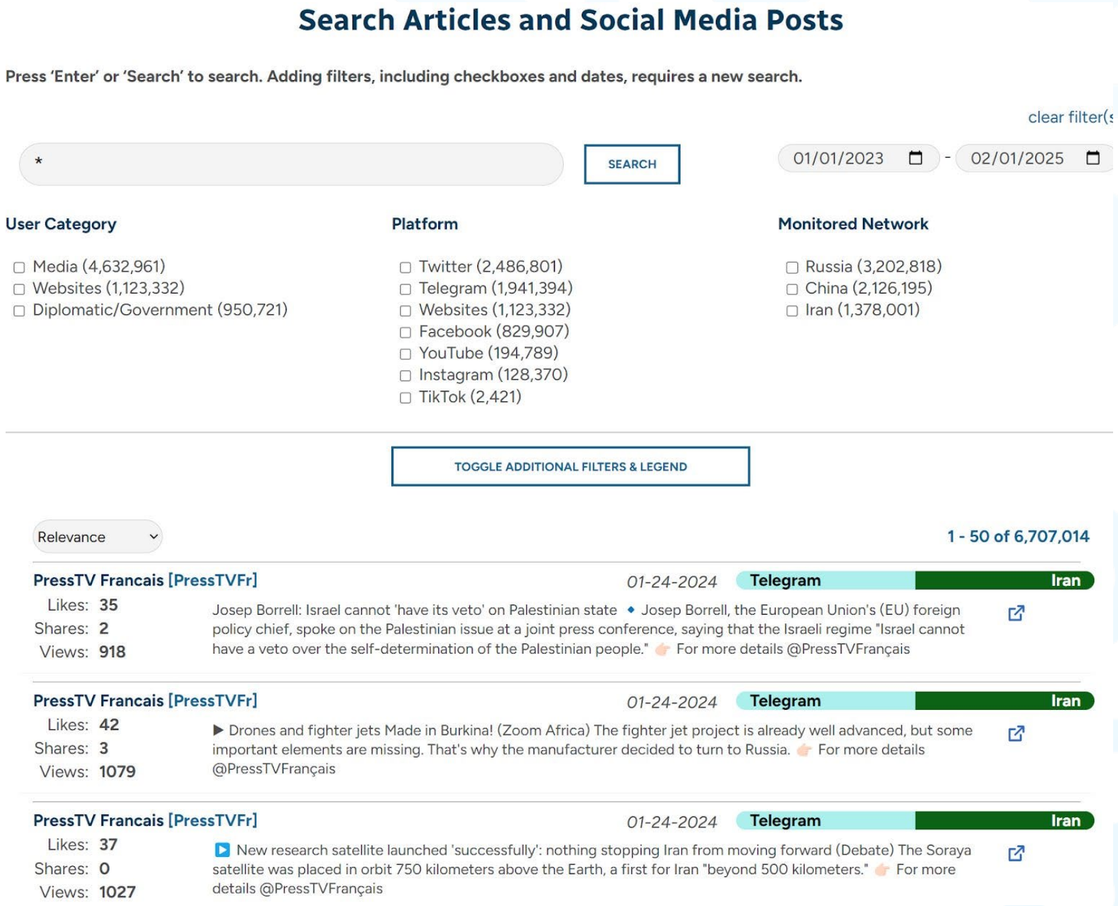
Tools for one or more social media platforms

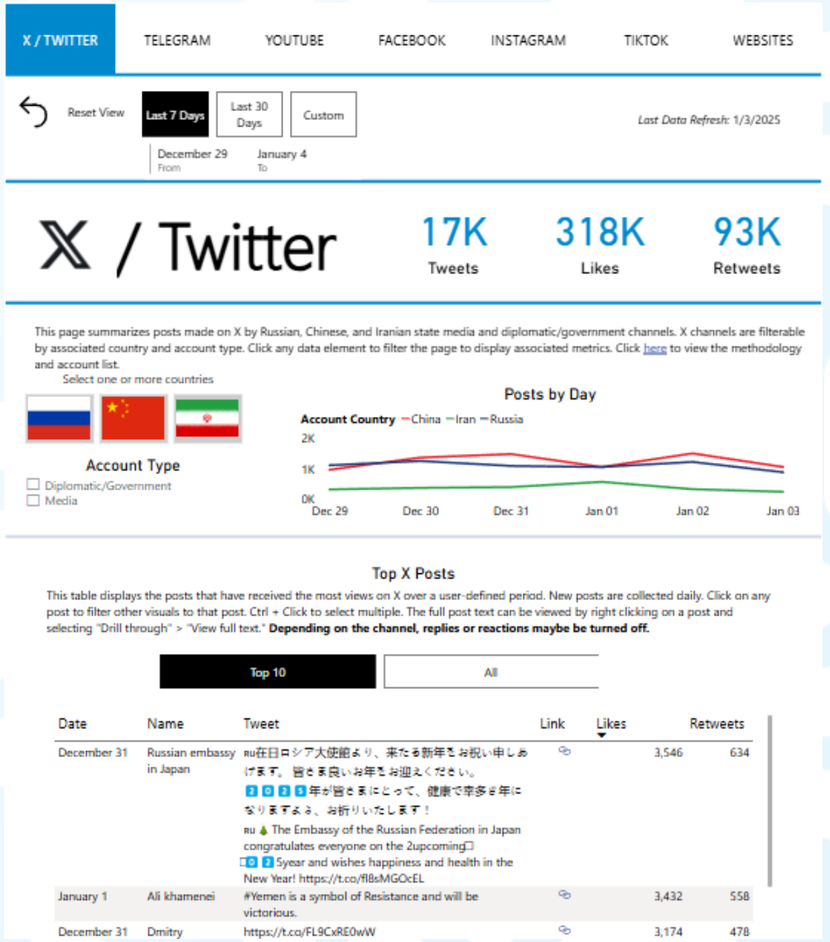
Facebook	Instagram
Telegram	Tiktok
Twitter/X	Youtube
Other Platforms	Multiple Platforms

Information Laundromat



Hamilton 2.0 Dashboard



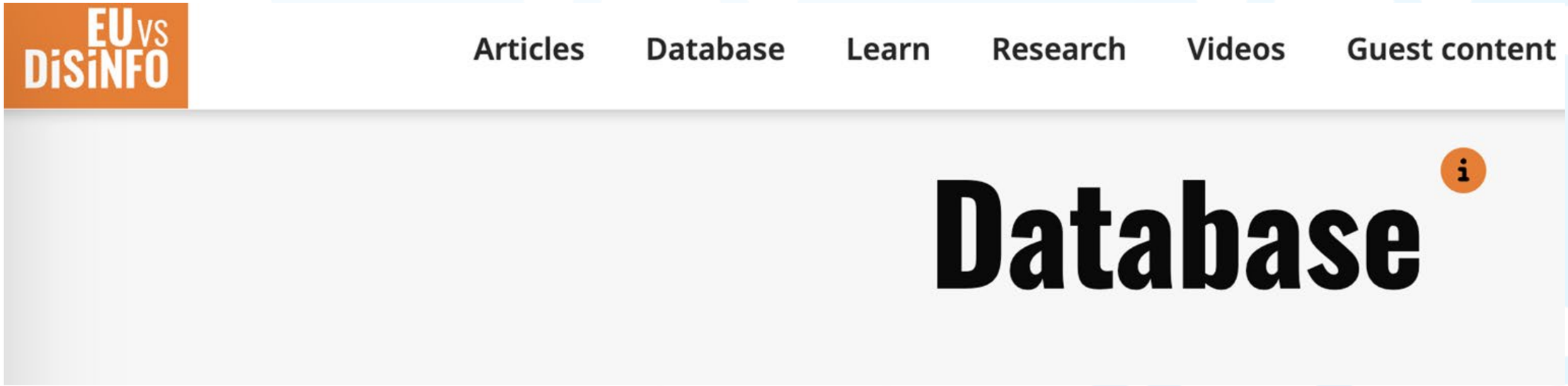


EU CyberNet

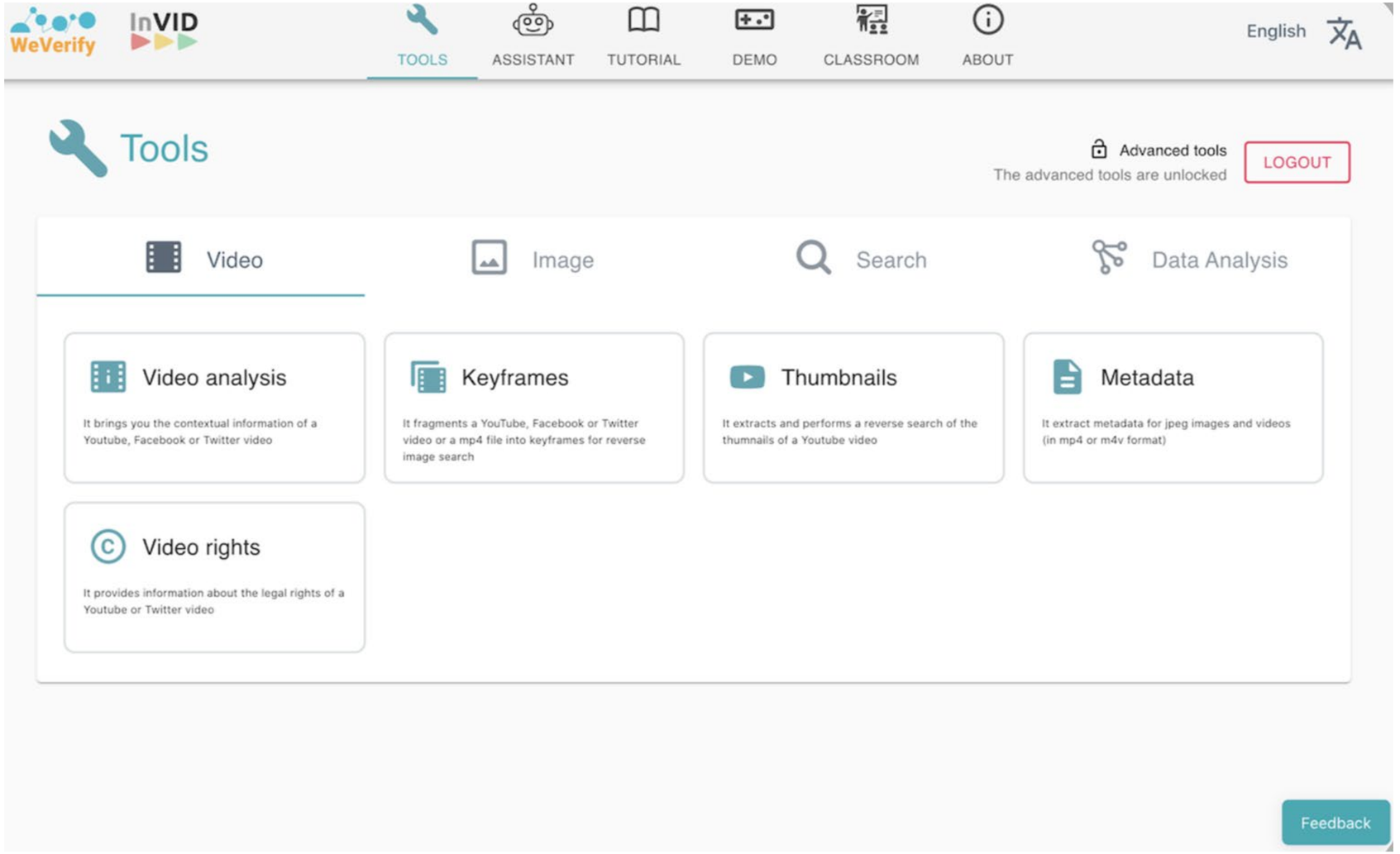
Federal Foreign Office

Funded by the European Union

EU vs Disinfo Database



WeVerify


WeVerify
InVID
TOOLS
ASSISTANT
TUTORIAL
DEMO
CLASSROOM
ABOUT
English
Tools
Advanced tools
The advanced tools are unlocked
LOGOUT
Video
Image
Search
Data Analysis
Video analysis
It brings you the contextual information of a Youtube, Facebook or Twitter video
Keyframes
It fragments a YouTube, Facebook or Twitter video or a mp4 file into keyframes for reverse image search
Thumbnails
It extracts and performs a reverse search of the thumnails of a Youtube video
Metadata
It extract metadata for jpeg images and videos (in mp4 or m4v format)
Video rights
It provides information about the legal rights of a Youtube or Twitter video
Feedback

Viginum D3lta

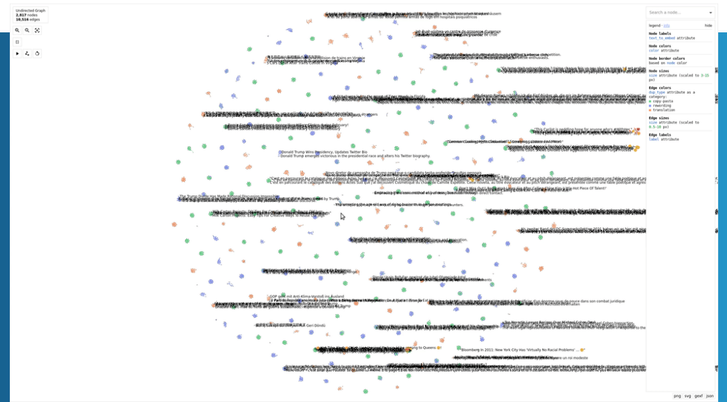
D3lta

If you like our project, please give us a star ⭐ on GitHub for the latest update.

Arxiv 2312.17338

This repository is the official implementation of D3lta, a library for detecting duplicate verbatim contents within a vast amount of documents.

It distinguishes 3 types of duplicate contents : copypasta (almost exact duplicates), rewording and translation. You can run it on CPU.



Commercial Detection Platforms

Social Listening Tools: Meltwater, Brandwatch, Junkopedia

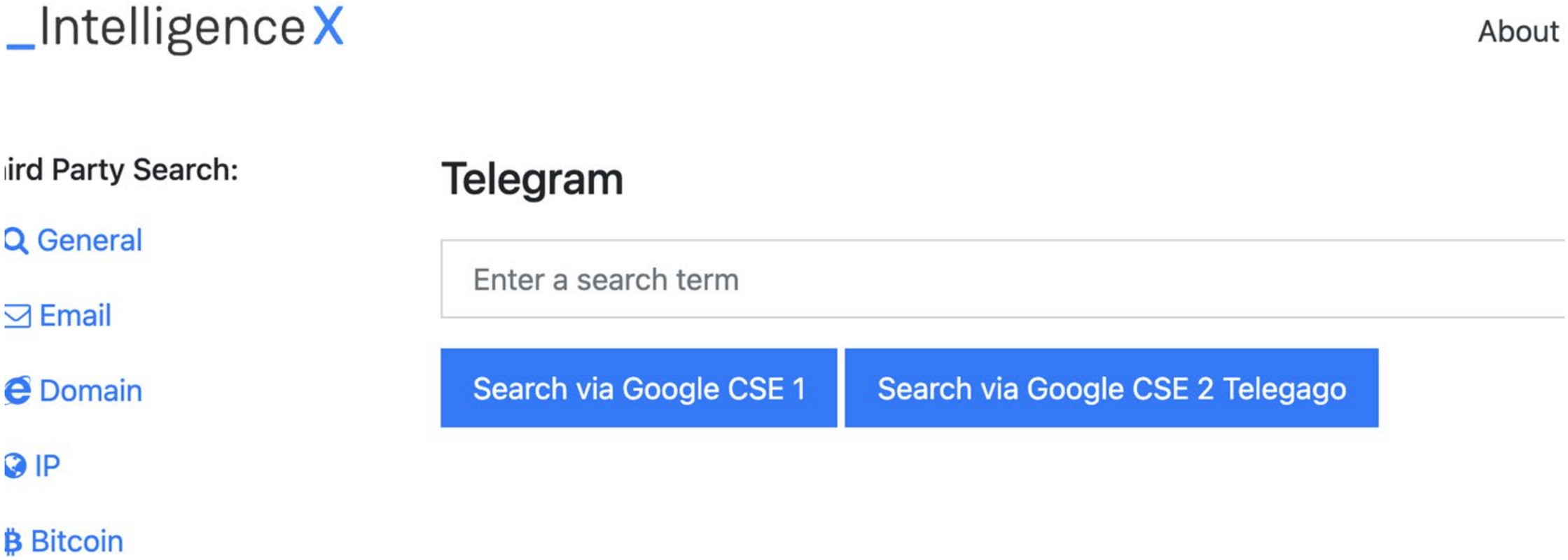
EU CYBERNET SUMMER SCHOOL 2025
Cyber Crisis Management: Navigating Disinformation and Cyber Attacks in the AI Era

EU CyberNet

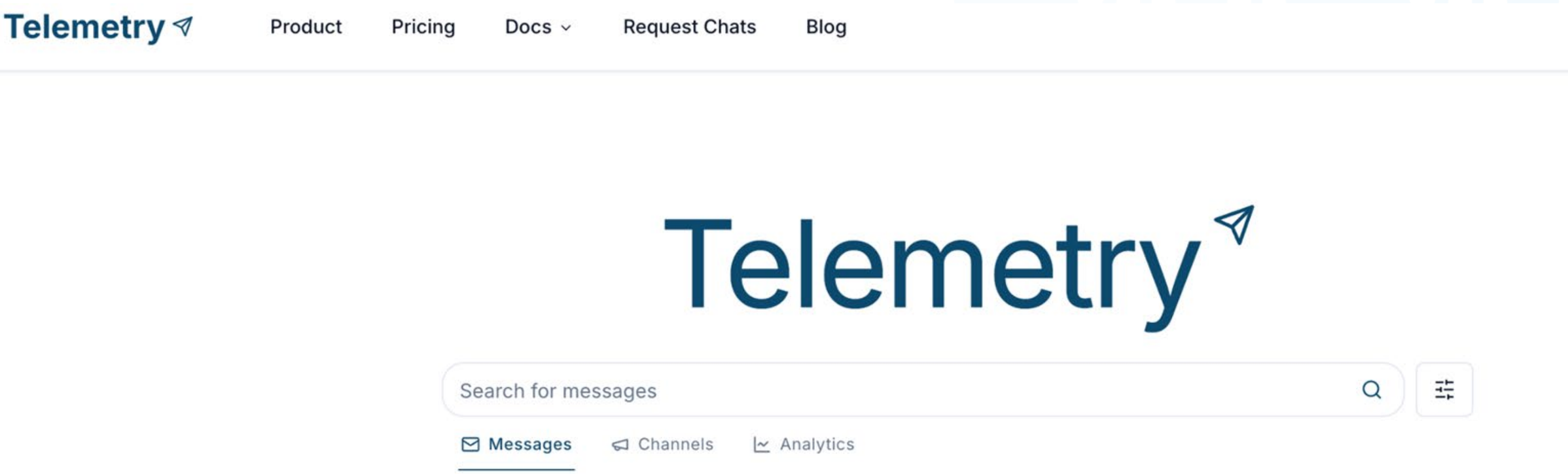
Federal Foreign Office

Funded by the European Union

IntelX


_IntelligenceX
About
ird Party Search:
General
Email
Domain
IP
Bitcoin
Telegram
Enter a search term
Search via Google CSE 1
Search via Google CSE 2 Telegago

Telemetry


Telemetry
Product
Pricing
Docs
Request Chats
Blog
Telemetry
Search for messages
Messages
Channels
Analytics

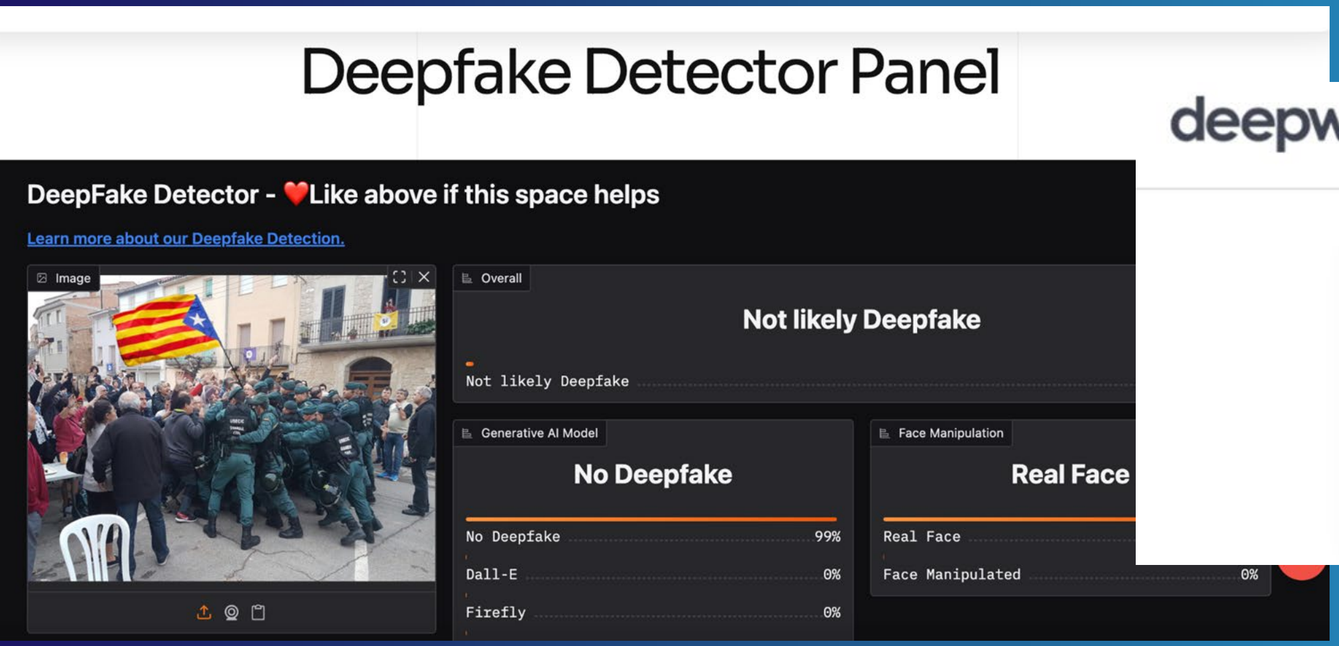
A number of tools on the market (this is not all)

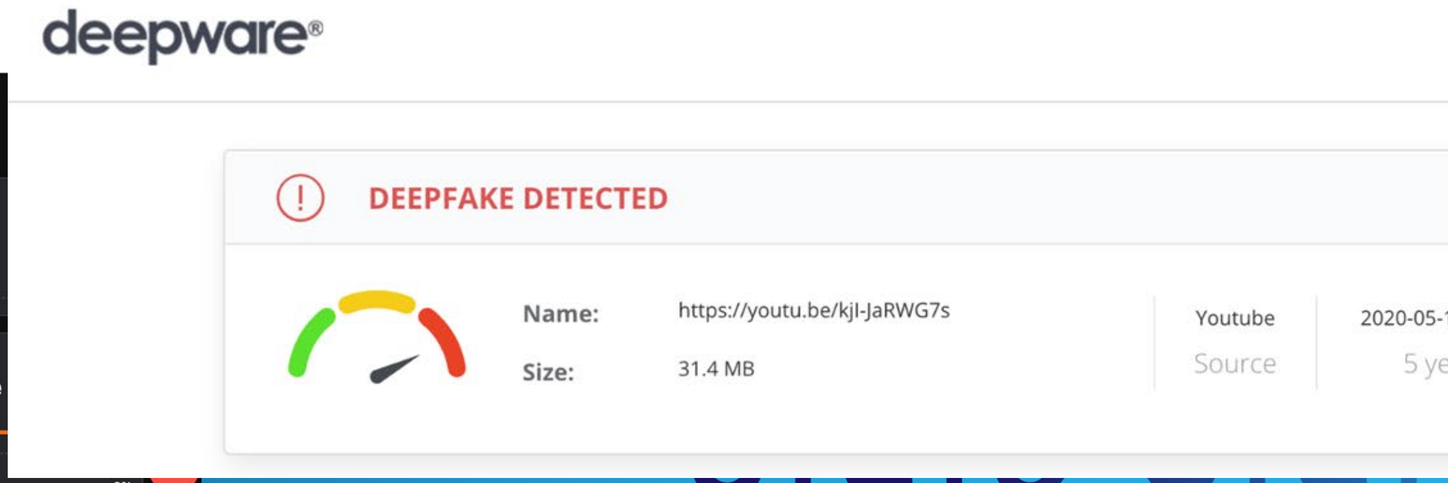
Things to note (tool tips)

EU CyberNet
Federal Foreign Office
Funded by the European Union

Most of the time deepfake detection requires the right tools

Audio for audio, Video for video: in a video with audio try to analyze separately





EU CyberNet
Federal Foreign Office
Funded by the European Union

Tools may not always be accurate

Check if image is AI-generated

We're quite confident that NO AI was used when producing this image.

RESET

Was this picture generated by ai or humans?

Try these examples: MidJourney | DALL-E | Stable Diffusion | Flux | Human

PDF Report

Prediction

real

This image is highly likely to be a real photo taken by a human.

AI Generated
Human Generated

0.02%
AI Probabilit

Upload a image

Analyzed Image

Analysis Results

77%

Authenticity Score

Analysis Breakdown

Metadata 85%

Analysis of image properties and format

Noise 48%

Evaluation of natural image noise patterns

Compression 94%

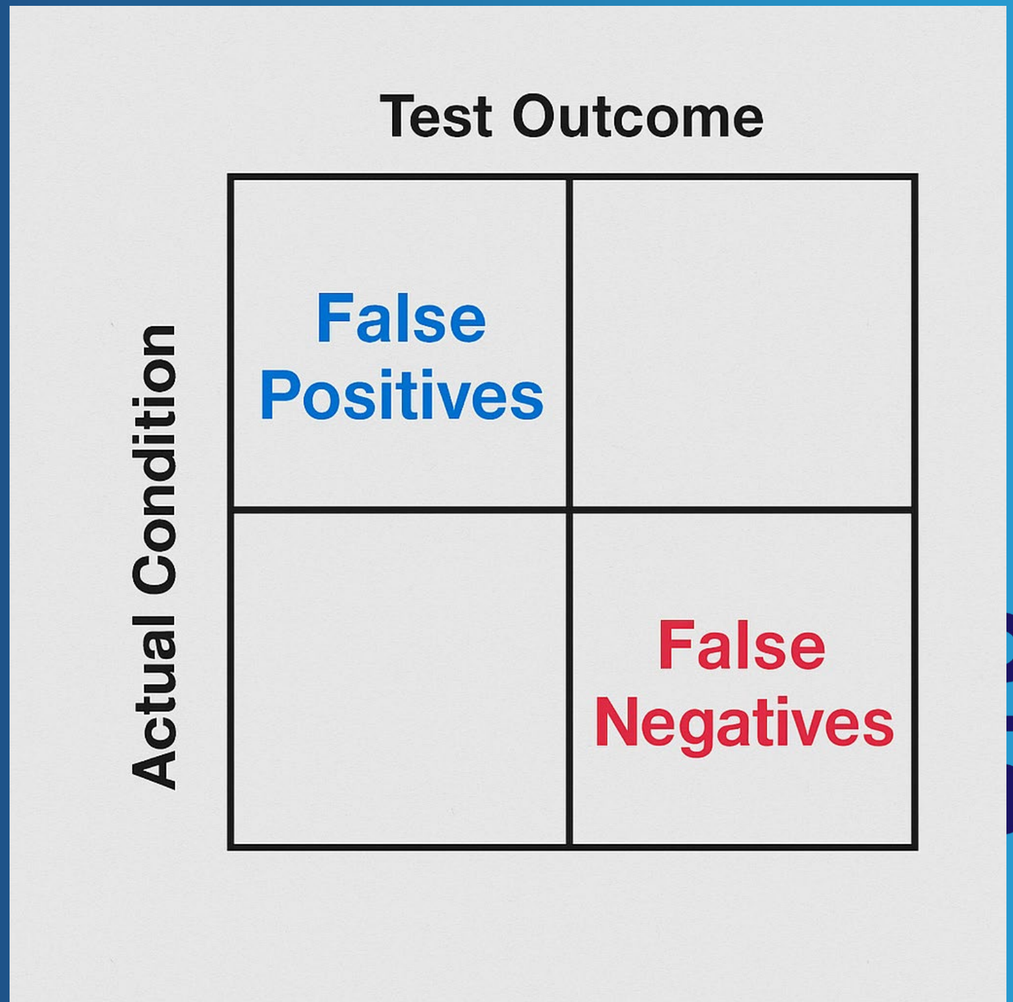
Assessment of compression artifacts

Authenticity 81%

Detection of artificial patterns

EU CyberNet
Federal Foreign Office
Funded by the European Union

You can get both false positives and false negatives



EU CyberNet

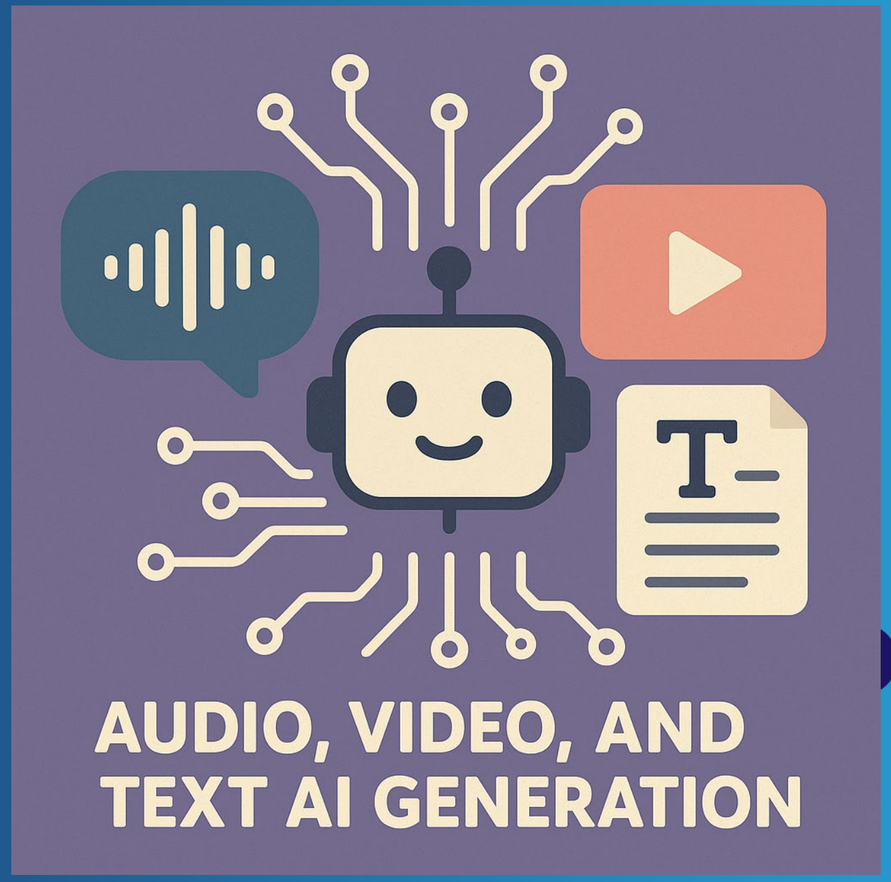
Federal Foreign Office

Funded by the European Union

Image quality matters

EU CyberNet
Federal Foreign Office
Funded by the European Union

Tools may not have caught up with new tech



EU CyberNet
Federal Foreign Office
Funded by the European Union

Provenance is often better than just detection

Look for the history of publication rather than a simple reading of if it was published or not

C2PA Coalition for Content Provenance and Authenticity

EU CyberNet

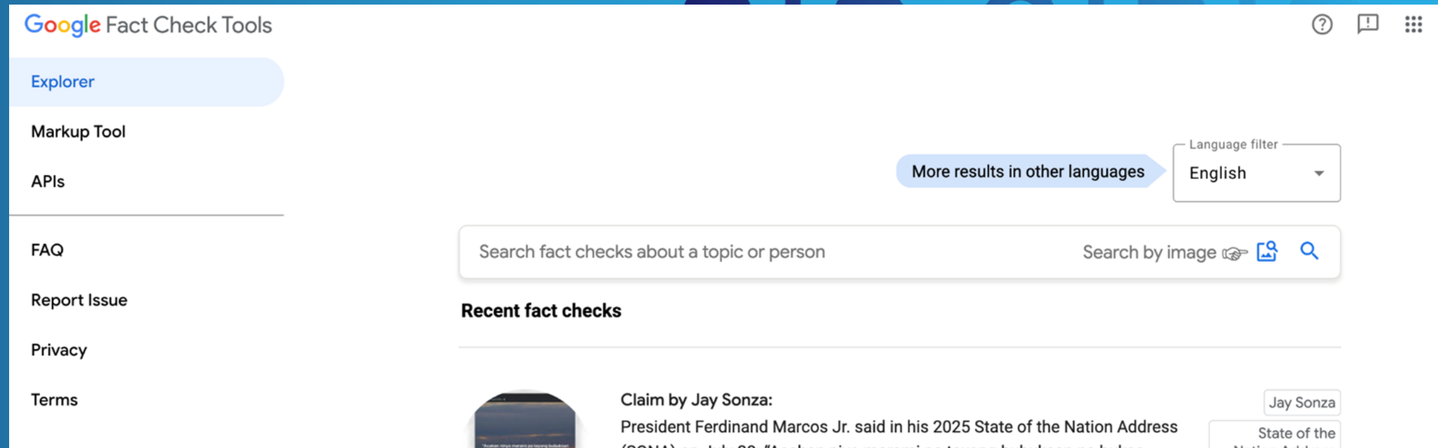
Federal Foreign Office

Funded by the European Union

Cross checking many tools and sources

TrueMedia.org

InVID

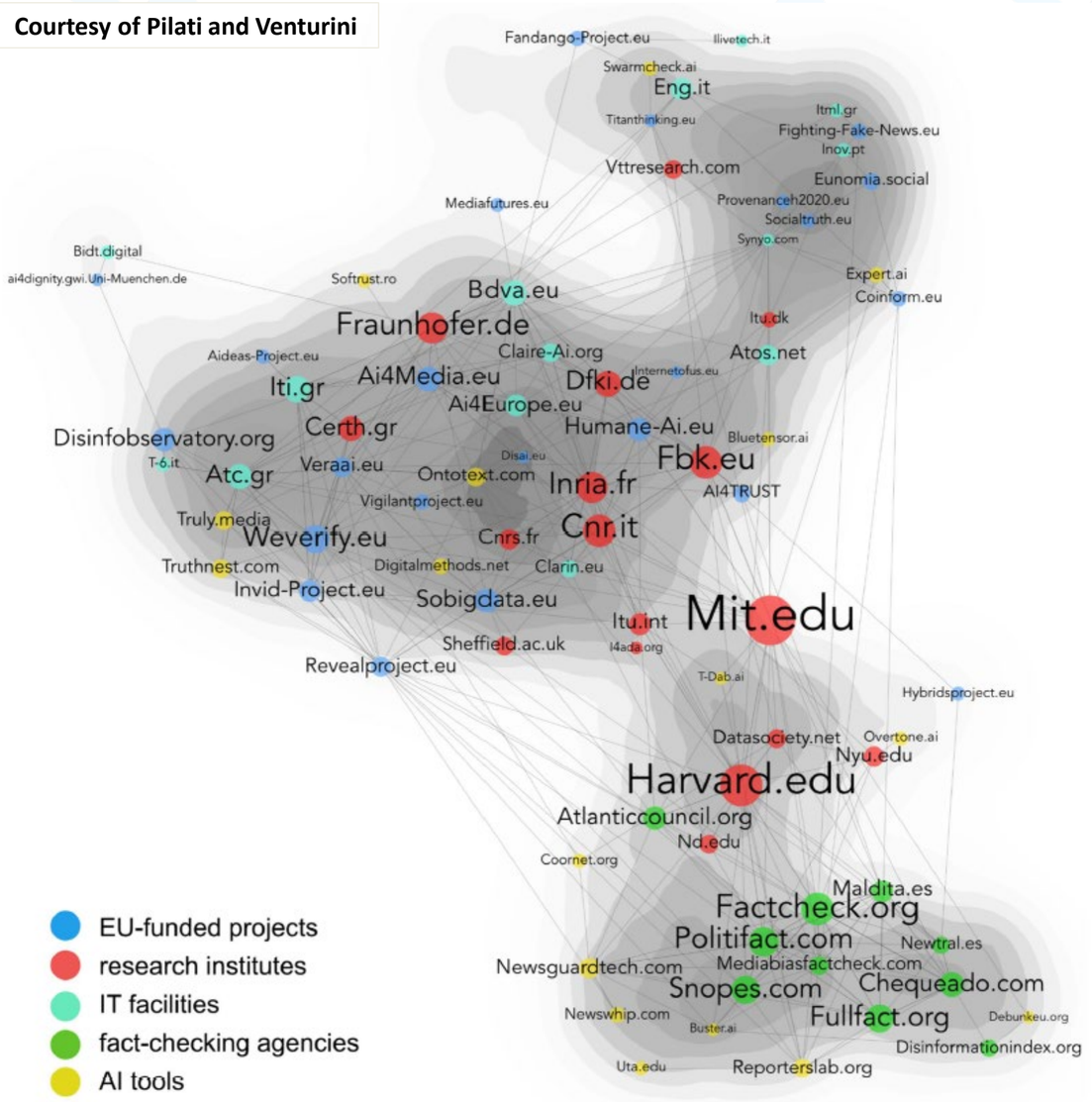


EU CyberNet
Federal Foreign Office
Funded by the European Union

BACKUP SLIDES

AI Initiatives Against Disinformation

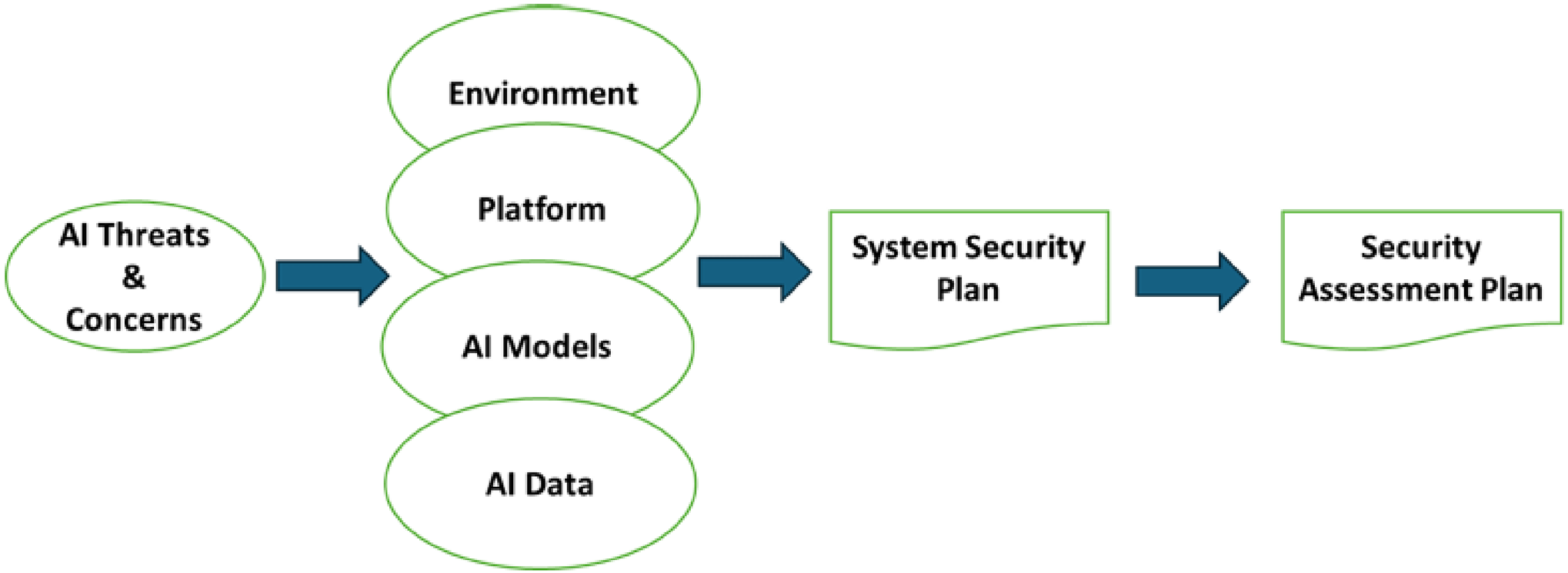
Courtesy of Pilati and Venturini



Courtesy of Goldstein et. al.

		Promise; if implemented...	Limitation
Model Design & Construction	AI Developers Build Models With More Detectable Outputs	Influence operations with language models will be easily discoverable	Technically challenging and requires coordination across developers
	AI Developers Build Models That Are More Fact-Sensitive	Language models will be less effective at spreading falsehoods	Technical methods are still being explored; may only impact some influence operations
	Developers Spread Radioactive Data to Make Generative Models Detectable	Makes it easier to detect if content is AI generated	Technically uncertain and may be easily circumvented
	Governments Impose Restrictions on Training Data Collection	Limits creation of new models (but only for those in jurisdictions that comply)	Data access restrictions would require high political will
	Governments Impose Access Controls on AI Hardware	Prevents some future models from being developed altogether	Restrictions on semiconductors could escalate geopolitical tensions and hurt legitimate businesses
Model Access	AI Providers Impose Stricter Usage Restrictions on Models	Makes it more difficult for propagandists to obtain cutting-edge models for campaigns	Requires coordination across AI providers and risks hurting legitimate applications
	AI Providers Develop New Norms Around Model Release	Restricts access to future models, but unlikely to prevent propagandists from obtaining already-public ones	Requires coordinating across AI providers and could concentrate capabilities among a small number of companies
	AI Providers Close Security Vulnerabilities	Prevents misuse and access of models via theft and tampering	Only affects one route to model access
Content Dissemination	Platforms and AI Providers Coordinate to Identify AI Content	Increases the likelihood of detecting AI-enabled influence operations	Will not affect platforms that do not engage; may not work in encrypted channels
	Platforms Require "Proof of Personhood" to Post	Increases the costs of waging influence operations	Current proof of personhood tests are often gameable by determined operators
	Entities That Rely on Public Input Take Steps to Reduce Their Exposure to Misleading AI Content	Protects entities relying on public inputs from AI-enabled campaigns	Significant changes to public comment systems could disincentivize participation
	Digital Provenance Standards Are Widely Adopted	Increases detection of AI-generated content	Significant changes would require large-scale coordination
Belief Formation	Institutions Engage In Media Literacy Campaigns	Mitigates the impact of influence operations	May reduce trust in legitimate content
	Developers Provide Consumer-Focused AI Tools	Increases the likelihood of people consuming high quality information	AI tools may be susceptible to bias; users may become overly reliant on them

NIST AI Threat Assessment

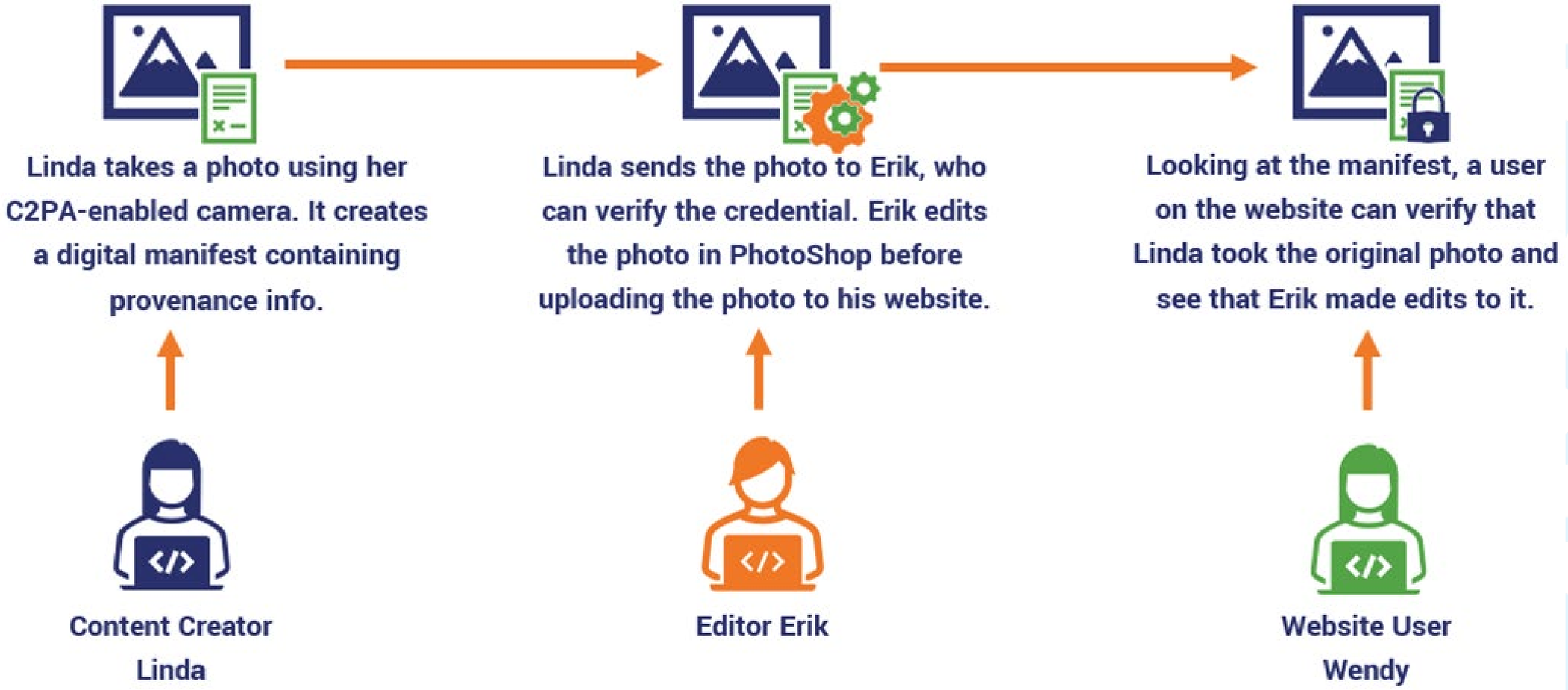


EU CyberNet
Federal Foreign Office
Funded by the European Union

NIST AI Risk Management Framework

Reconnaissance	Resource Development	Initial Access	ML Model Access	Execution	Persistence	Privilege Escalation	Defense Evasion	Credential Access	Discovery	Collection	ML Attack Staging	Exfiltration	Impact
AML.TA0002	AML.TA0003	AML.TA0004	AML.TA0000	AML.TA0005	AML.TA0006	AML.TA0012	AML.TA0007	AML.TA0013	AML.TA0008	AML.TA0009	AML.TA0001	AML.TA0010	AML.TA0011
AML.T0006 & Active Scanning	AML.T0008 Acquire Infrastructure	AML.T0015 Evade ML Model	AML.T0040 AI Model Inference API Access	AML.T0050 & Command and Scripting Interpreter	AML.T0018 Backdoor ML Model	AML.T0054 LLM Jailbreak	AML.T0015 Evade ML Model	AML.T0055 & Unsecured Credentials	AML.T0063 Discover AI Model Outputs	AML.T0036 & Data from Information Repositories	AML.T0018 Backdoor ML Model	AML.T0025 Exfiltration via Cyber Means	AML.T0034 Cost Harvesting
AML.T0004 Search Application Repositories	AML.T0002 Acquire Public ML Artifacts	AML.T0049 & Exploit Public-Facing Application	AML.T0044 Full ML Model Access	AML.T0053 LLM Plugin Compromise	AML.T0051 LLM Prompt Injection	AML.T0053 LLM Plugin Compromise	AML.T0054 LLM Jailbreak		AML.T0062 Discover LLM Hallucinations	AML.T0037 & Data from Local System	AML.T0043 Craft Adversarial Data	AML.T0024 Exfiltration via ML Inference API	AML.T0029 Denial of ML Service
AML.T0001 Search for Publicly Available Adversarial Vulnerability Analysis	AML.T0017 & Develop Capabilities	AML.T0051 LLM Prompt Injection	AML.T0047 ML-Enabled Product or Service	AML.T0011 & User Execution	AML.T0061 LLM Prompt Self-Replication	AML.T0051 LLM Prompt Injection	AML.T0051 LLM Prompt Injection		AML.T0007 Discover ML Artifacts	AML.T0035 ML Artifact Collection	AML.T0005 Create Proxy ML Model	AML.T0057 LLM Data Leakage	AML.T0059 Erode Dataset Integrity
AML.T0000 Search for Victim's Publicly Available Research Materials	AML.T0021 & Establish Accounts	AML.T0010 ML Supply Chain Compromise	AML.T0041 Physical Environment Access		AML.T0020 Poison Training Data				AML.T0014 Discover ML Model Family		AML.T0042 Verify Attack	AML.T0056 LLM Meta Prompt Extraction	AML.T0031 Erode ML Model Integrity
AML.T0003 Search Victim-Owned Websites	AML.T0016 & Obtain Capabilities	AML.T0052 & Phishing							AML.T0013 Discover ML Model Ontology				AML.T0015 Evade ML Model
	AML.T0020 Poison Training Data	AML.T0012 & Valid Accounts							AML.T0056 LLM Meta Prompt Extraction				AML.T0048 External Harms

Content Provenance Initiative



Data Provenance Standards