Coordinated Inauthentic Behaviour Detection Tree

Romero Vicente, Ana
Researcher, EU DisinfoLab
16/06/2025

This document part of the veraAI project, to be included in Deliverable D4.2

Table of content

Definition of CIB

EU regulation

- How CoP, DSA and EDAP understand CIB
- There is not an accurate definition

Platforms

- CIB reported in the policies of different social media platforms
- CIB definition differs from platform to platform

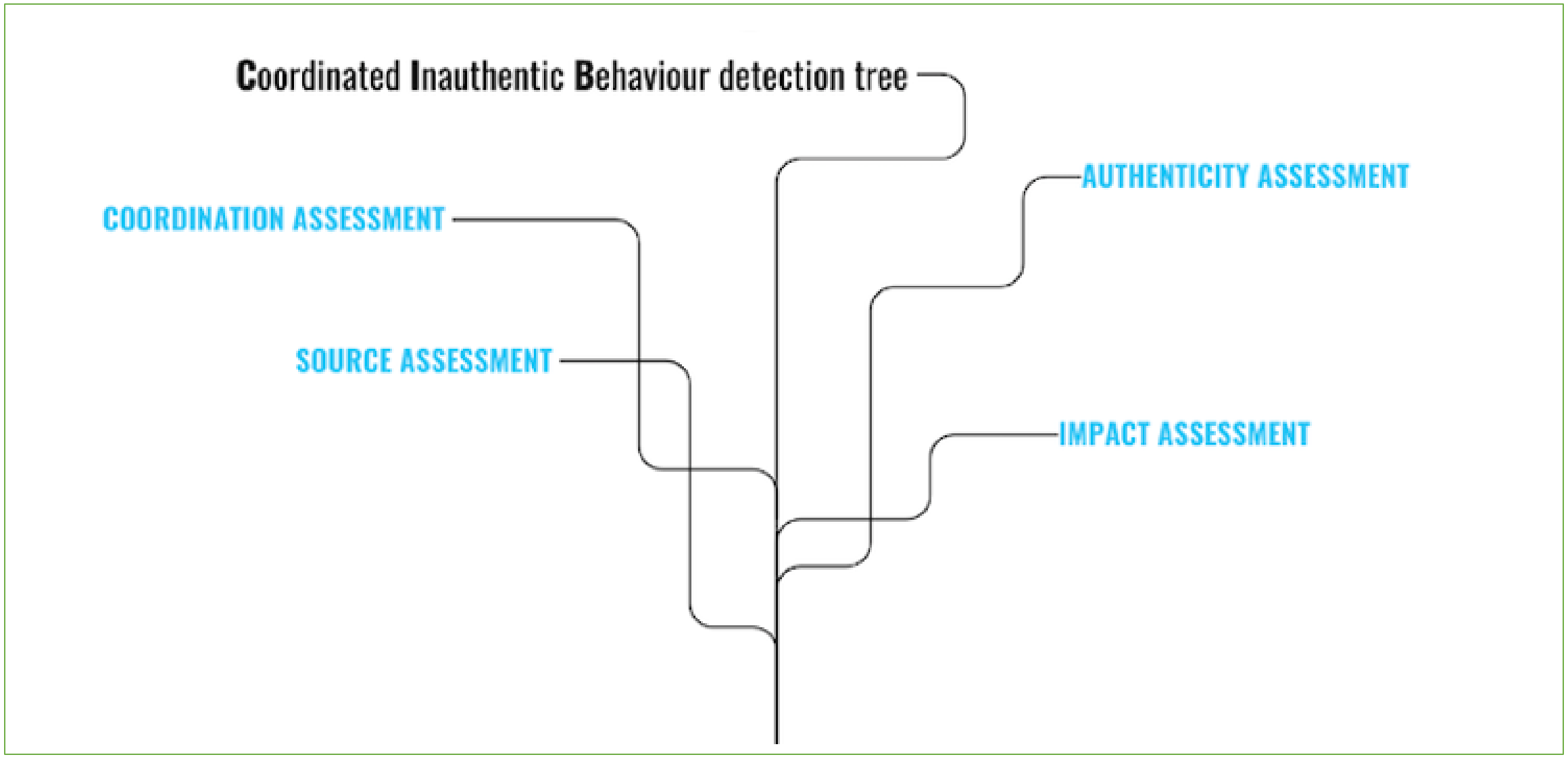
\+ *Absence of specific platform mechanisms to report CIB*

EU DisinfoLab approach

CIB is a manipulative communication strategy based on:

- Distribution and amplification of content in a coordinated and non-organic way
- Using mostly fake accounts, and sometimes even authentic accounts
- Circulation of content is fully, partially automated or not automated at all

CIB detection tree



Coordination assessment branch

How do we detect coordination?

- Structured arrangement of tasks and efforts
- Multiple agents collaborating (Coordination ≠ automatisation)
- Same goal - to deceive

All of this is compounded by the challenge posed by AI

Indicators of coordination

BEHAVIOURAL ANALYSIS

- Accounts created around the same time
- Similar posting timestamps
- Sudden spikes in messaging
- Significant activity from locations/ time zones

NETWORK ANALYSIS

- Accounts interacting synchronously
- Tightly interconnected clusters of accounts.
- Similar posting cross-platform patterns

CONTENT ANALYSIS

- Identical or similar textual content (in different languages too), hashtags, links, posts, images, videos or memes
- Single-topic accounts

IDENTITY ANALYSIS

- Same or similar profile name or bio

METADATA ANALYSIS

- Multiple accounts using the same IP address, device and configurations.

VISUAL ANALYSIS

- Same or similar profile picture or cover photo

Authenticity assessment branch

CIB pillar: assess the genuineness and legitimacy of the campaign's objectives, actors, and content
This is further complicated by the AI challenge.

CONTENT-LANGUAGE ANALYSIS

- Poor translation, misspelling, typos.
- Sense of unnaturalness/ repetitive tone
- Extreme content polarisation (amplifying specific narratives) or whimsical and conspiratorial plots
- Manipulated, forged or fabricated texts, including using fake endorsements by public figures.

NETWORK ANALYSIS

- Accounts with little activity that rarely interact with users outside of their network.
- High interaction posts from low-activity accounts with incomplete profiles or minimal content history.
- Unusual high levels of likes, shares, or comments.
- Abnormal changes in follower count over short periods of time

Indicators of inauthenticity

BEHAVIOURAL ANALYSIS

- Account activity suddenly resumed after a long period (dormant accounts).
- Proof that account has been hijacked for the campaign.

IDENTITY ANALYSIS

- Lack of personalisation: generic profiles with minimal personal information.
- Profile names with random strings of letters and numbers

AUTOMATION ANALYSIS

- Bot behaviour - unusually rapid engagement

VISUALS ANALYSIS

- Visual theft
- Simultaneous profile pictures updates across multiple accounts
- Spoofed, fabricated AI-generated visuals

METADATA ANALYSIS

- Activity from IP ranges associated with VPNs or proxies
- Sudden spikes in API requests that far exceed normal patterns
- Some IP geographic locations could indicate bot farms

Source assessment branch

- Attribution is challenging and difficult:
    - Lack of data access
    - The AI challenge
- Tracking the source requires a multi-faceted approach

Source tracking

NETWORK ANALYSIS

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

METADATA ANALYSIS

- Monitor IP addresses, user agents, timestamps, request details
- Track the geographical distribution of API requests to identify physical locations
- Metadata analysis of images, videos, links

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 (faces, places)

Distribution and Impact assessment branch

- It is not only about impact but also **distribution**
- The main purpose of a CIB network is to **amplify content mainly through fake means.** AI-based technologies represent a challenge in this field.
- **Exploring the distribution** helps learning more about the CIB and **its impact**

Distribution and impact tracking

BEHAVIOURAL ANALYSIS

- Peripheral accounts amplifying the content of the core account(s)
- Accounts having specific roles in the amplification process.

CONTENT ANALYSIS

- Extreme content polarisation
- Evaluate if campaign content targets a global audience or specific regions to gauge its reach

NETWORK ANALYSIS

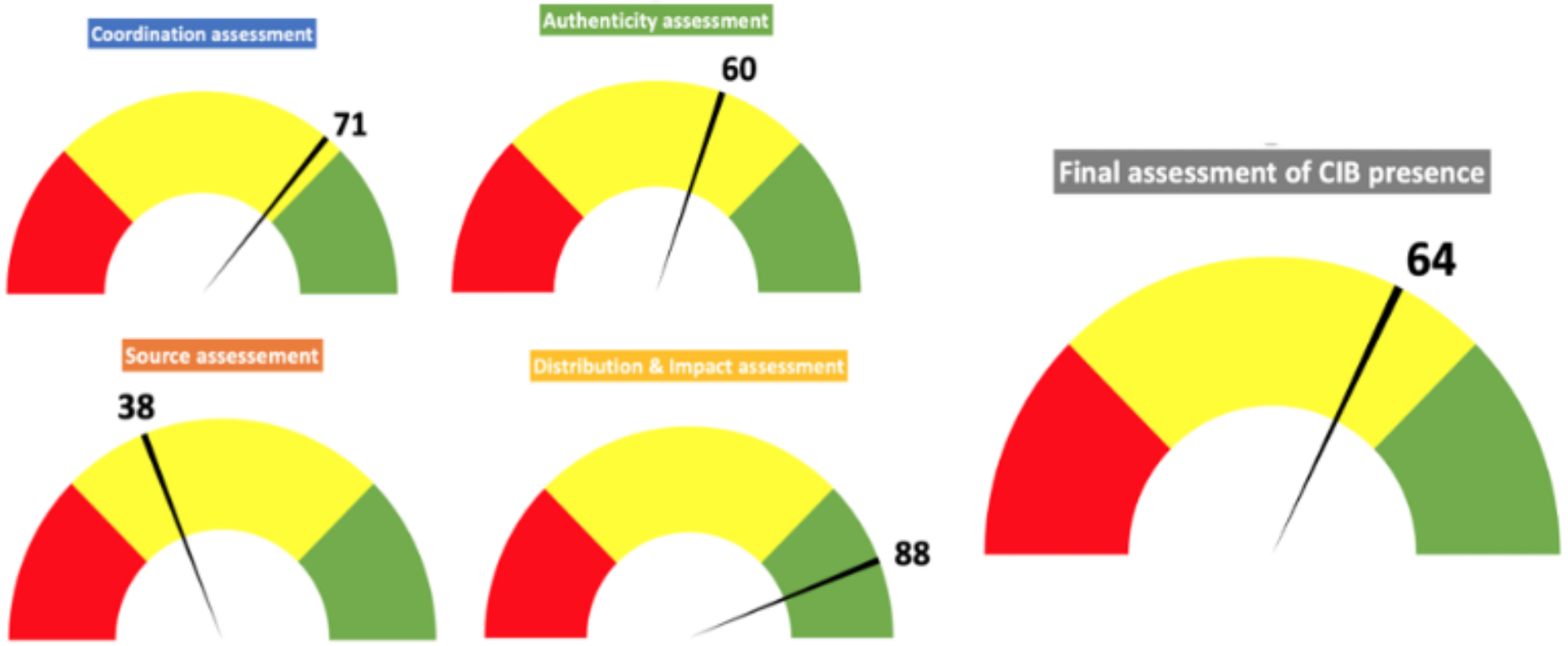
- Unusual volumes of likes or reshares
- Media amplification
- Public figures/influencers amplification
- Various methods from sentiment analysis to tracking trending hashtags
- Track backlinks
- Use multiple social media platforms to maximise outreach
- Only certain viewpoints are amplified. Dissenting opinions

From framework to fieldwork: Visual CIB assessment

- 50 indicators across branches coordination, authenticity, source, impact
- Generates a CIB likelihood score (0–100%)
- Results shown via 5 colour-coded gauges
- Designed for clarity and non-expert use
- Methodology applied to three published disinfo investigations:

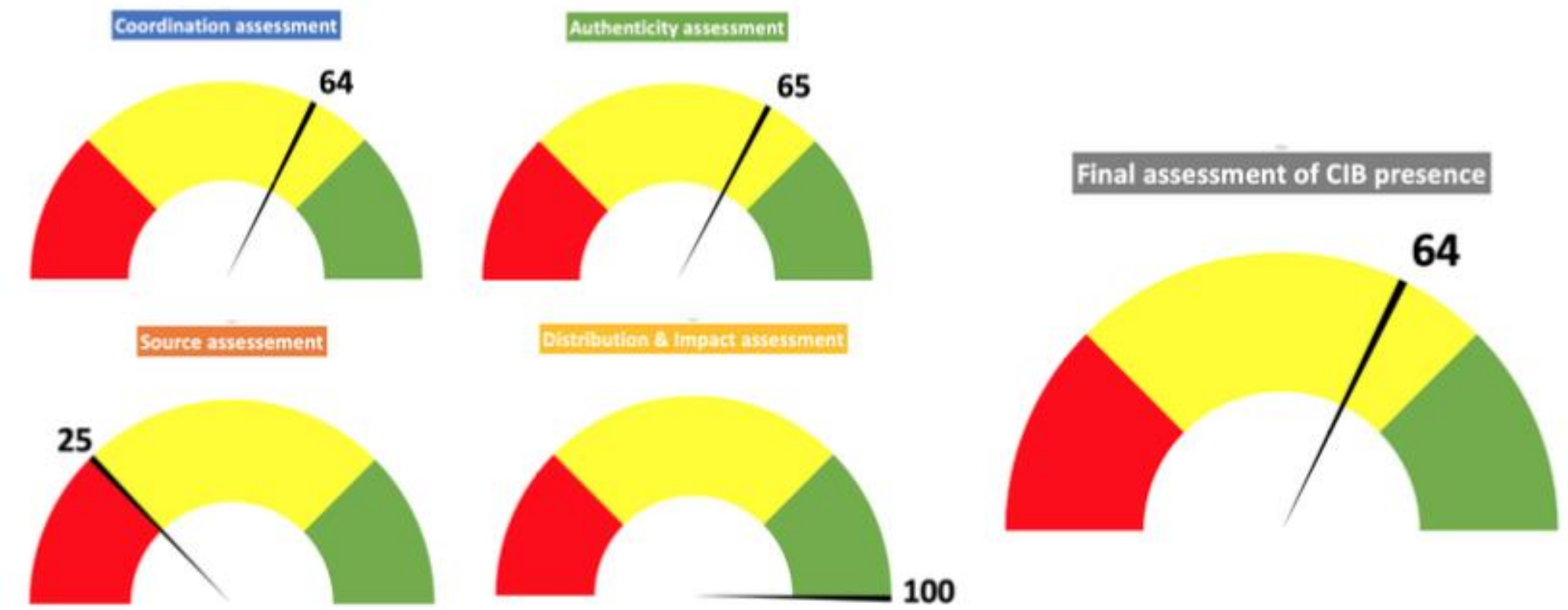
Case study 1

CheckFirst's Operation Overload Medium-high likelihood of CIB



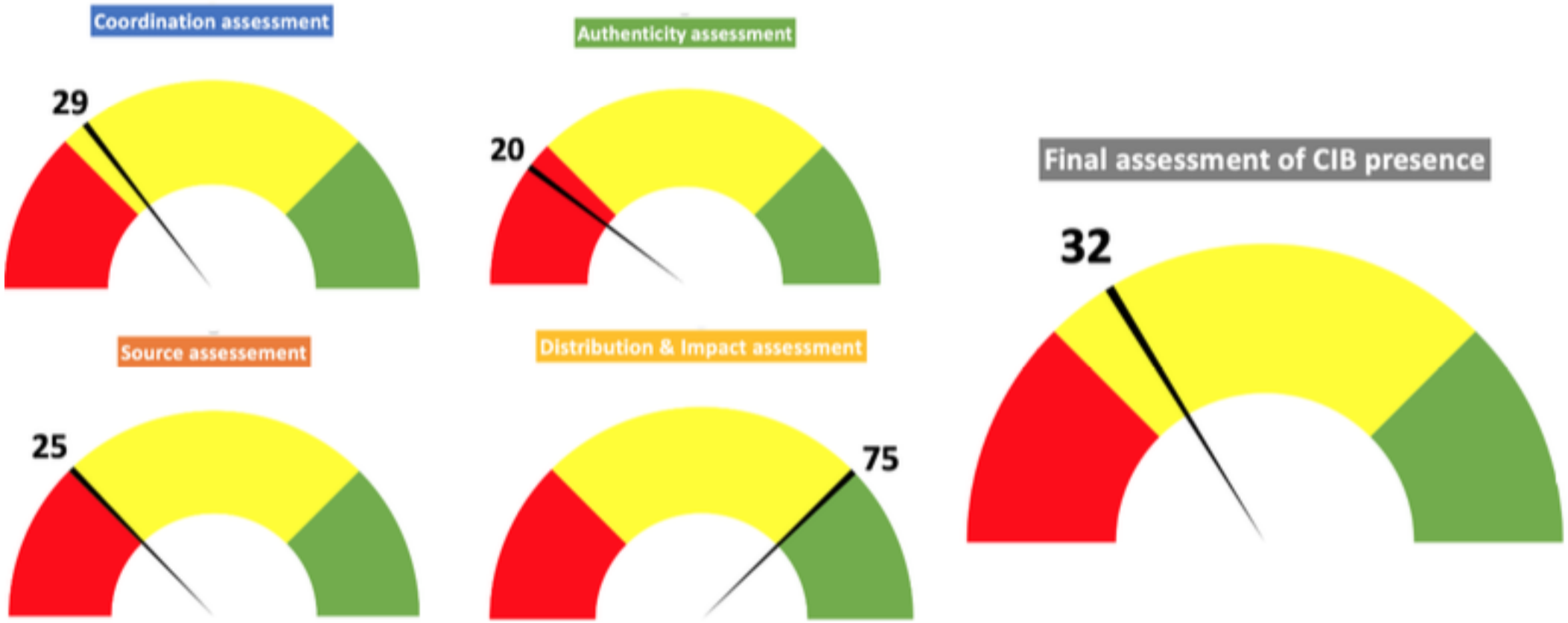
Case study 2

DFRLab & BBC Verify's TikTok Influence Operation **Medium-high likelihood of CIB**



Case study 3

QAnon's "Save the Children" campaign **Medium-low likelihood of CIB**



Conclusions

- CIB is widely recognized but…
    - Definition is still left up to individual platforms
    - The EU regulation lacks a specific one
    - EU DisinfloLab proposes a comprehensive, generally applicable definition
- Absence of specific platform mechanisms to report CIB
- There is not a single feature that can definitively prove CIB, but a combination of several indicators to be carefully evaluated
- AI: enemy & ally
- Our CIB visual assessment offers a transparent framework for identifying likely CIB and bridges the gap between qualitative analysis and quantifiable evidence.
- We **welcome feedback** and suggestions to refine the **CIB detection tree** and **CIB vsual assessment**

Thank you

Do you have any questions?

Ana Romero Vicente / Researcher at EU DisinfoLab

Contact: ar@disinfo.eu

vera.ai

CERTH CENTRE FOR RESEARCH & TECHNOLOGY HELLAS

The University Of Sheffield.

ATC ATHENS TECHNOLOGY CENTER

KINIT

ontotext

Deutsche Welle

Università degli Studi di Napoli Federico II

1506 Università degli Studi di Urbino Carlo Bo

AFP

Universiteit van Amsterdam

école normale supérieure paris-saclay

université PARIS-SACLAY

EBU OPERATING EUROVISION AND EURORADIO

EU DisinfoLab

Fraunhofer IDMT

Follow us on Twitter: @veraai_eu

Website: https://www.veraai.eu/

Co-financed by the European Union, Horizon Europe programme, Grant Agreement No 101070093.

Additional funding from Innovate UK grant No 10039055 and the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract No 22.00245