

OSINT

Resource · 2026 edition · 60 min read

THE CIB DETECTION TREE: 2026 EDITION

by Antoine de Gunzbourg

CHAPTER 1/5 · CIBS IN THE WORDS OF THE PLATFORMS

How platforms define Coordinated Inauthentic Behaviour, should detection start with narrative, with behaviour or with a mix of both, and what AI changes for attackers and defenders.

Méléta Académie · October 2026 · meleta.eu/ressources/cib-detection-tree-2026-platforms

CONTENTS

Introduction

What are CIBs?

Narrative vs Behaviour

The disappearance of the “content / behaviour” separation

Are CIBs Political or Commercial?

Why a detection tree

Detection methodology

Definitions of CIB by the platforms

Facebook

Instagram

X (formerly Twitter)

Bluesky

LinkedIn

WhatsApp

YouTube

TikTok

Discord

Amazon

CIBs and AI

Conclusion

INTRODUCTION

The original version of the CIB detection tree **was published in 2021**. The idea was to conceive a methodology for detecting and proving Coordinated Inauthentic Behaviours across platforms. The tree was a list of detection techniques employed by analysts to detect or prove Coordinated Inauthentic Behaviour, organised in four branches: The Source Assessment to investigate who the accounts were, the Coordination Assessment to investigate the links between the accounts and known entities, the Impact Assessment to investigate the metrics, as metrics inflation is a big part of CIBs, and the Authenticity Assessment to detect or to prove inauthenticity of the accounts.

This update from the previous 2021 CIB Detection Tree will check new techniques, check if the old techniques are still efficient, update the examples with more recent reports and CIB cases, review the new policies of the platforms and address the question of AI.

The publications will follow the initial tree branches and will be released progressively, branch by branch.

To the four initial branches we are adding a fifth part, a chapter dedicated to the study of the platforms' policies.

This chapter thus consists in reading, understanding and analysing the policies of different platforms: Facebook, Instagram, X, Bluesky, LinkedIn, WhatsApp, YouTube, TikTok, Discord, Amazon. We will be able to add more platforms in the future.

WHAT ARE CIBS?

CIB was a notion initially developed by Facebook. Their initial idea was to strictly separate policies based on behaviours and policies based on content. Therefore, their systems would be able to detect *“inauthentic behaviour”* without looking at and judging the content of the speeches. Content was covered by another set of policies: terms and conditions (including on hate speech, harassment etc.).

Initially, Facebook tried to base the definition of CIBs on the notion of *“Fake Accounts”*. But the idea of a fake account is not equally shared across platforms. Other platforms also adopted policies targeting the *“behaviours”* and used various names: platform manipulation or inauthentic behaviours.

Researchers all adopted the term CIB to characterise their discoveries in this area.

CIB appeared to be the best term because it was tying together three elements:

Coordinated: It has to be a network of accounts, working together with a shared agenda.

Inauthentic: there have to be markers of inauthenticity and manipulation attempts. The *“inauthentic”* criterion is also useful to bypass the *“fake / true”* problem that was forcing researchers to prove something was fake.

Behaviour: Behaviours are measurable and tangible. They can be detected by Boolean logic: account creation date, time of posting, profile picture, language of posting: there is no need to interpret content or decide between opinion and fact.

But essentially, CIBs are different for platforms and for researchers. In a word:

For researchers, CIB means investigation.

For platforms, CIB means spam.

For researchers, CIBs are networks of inauthentic accounts trying to amplify an adversarial narrative or an Influence Operation. Finding a CIB network can help prove the inauthenticity or the malevolent intention of a campaign, or help to attribute an influence operation to a known actor. Finding a CIB means investigating it, preserving the evidence and tracing it back to attribution.

For platforms, detecting inauthentic behaviours was mostly an upgrade of their anti-spam policies. In 2011, Facebook was conceiving a *“Facebook Immune System”* (css.csail.mit.edu ) , an anti-spam policy that was already trying to distinguish *“fake accounts”* and *“compromised accounts”* from authentic accounts. They were already trying to detect Fake Accounts created *“to friend or like objects en masse and thereby boost the reputation or ranking of those objects”* by using behavioural signals like the geolocation, the age of the account, post deletion and friend request rejections.

Platforms try to detect CIBs by automated tools, which leads to many false positives and failures to really detect CIBs. But it opened up the possibility of differentiating content moderation from behaviour-based methodologies.

NARRATIVE VS BEHAVIOUR

For both counter-disinformation organisations and platforms, the fight against CIB opened a previously unknown opportunity: to detect, prove and counter disinformation and influence operations while relying on purely behavioural or technical signals, and to avoid being trapped in studying narratives and content moderation.

First, relying on behavioural signals was proving more efficient for standardisation, knowledge sharing and to establish databases of known assets (like OpenCTI), a critical infrastructure in any CIB investigation or countering policy. Behavioural signals can more or less be standardised and shared among analysts (geolocation, infrastructure of the network, metrics such as number of likes or frequency of comments, timestamps of posting) and they are neutral.

The second appeal for “*behaviour*” methodologies was, for analysts and platforms alike, to avoid working on the content and the narratives. Moderation of content is a nightmare for platforms, which have to manually review every comment and have humans decide if content is harmful, violent, promotes terrorism or war crime apologies, hate speech or self-harm. Even worse: different societies and different countries with different regimes have different views and regulation of what content should be banned. Banning content on LGBTQ rights will reward you with a PR storm in one country and government applause in another one. TikTok for instance has this exact problem while trying to define what is a “*State affiliated media*”:

“The input we received (...) emphasized the importance of considering diverse political, legal and cultural contexts related to news media. In addition, we received feedback that there's no one-size-fits-all approach to state-affiliated media and supporting user education about the different ways that states seek to influence news and reporting is increasingly important.”

(TikTok, “TikTok's state-affiliated media policy”, 18 January 2023)

Although platforms try to exempt themselves from responsibility for the content published through their services (with the famous Section 230 of the Communications Decency Act, 47 U.S.C. § 230 (1996): [law.cornell.edu](https://www.law.cornell.edu/ucc/47usc/47usc230.html) > for instance), different layers of international regulations and policies, the latest being the Digital Services Act, have imposed on them a certain degree of responsibility for the content they publish.

Behaviour-based methodology allowed the platforms to fight against malevolent actors and to rely purely on detecting behavioural and technical signals without having to look at the content, involve human moderation and engage in heated and complicated debates for them on whether homosexuality is against nature, war crimes are just if the war is right, or BBC affiliation to the UK government equals Russia Today affiliation to the Kremlin. Even more appealing for them: they could upgrade their anti-spam systems and train them to detect technical and behavioural signatures.

For disinformation analysts too, relying on behavioural signals and *“content agnostic”* methodologies was a promise of neutrality, after a decade of being called *“arbiter of truth”*. It was also a newly found legitimacy to counter Influence Operations. Secret European programmes and NGOs tracking down EU citizens expressing *“pro-russian”* sentiment would be largely inadmissible, even if they tried to hide behind using the label *“Foreign”* on the influence operations. The *“Foreign”* label remains difficult to attribute, and has a long history of being abused in political contexts, as anyone can see in the use of the expression *“Foreign agent”* or *“party of foreign interests”*.

Behaviour-based methodologies were the promise of avoiding the narrative trap: having to constantly dance on the edge of freedom of expression and freedom of opinion, and having to analyse in depth nonsensical conspiracy theories or layers and sublayers of authoritarian propaganda.

Therefore, platforms and anti-disinformation organisations alike hoped to rely purely and solely on behavioural signals and content-agnostic methodologies to find CIBs. This came to a degree where some platforms tried to strictly separate teams and departments, one for CIB studying behaviours and one for content moderation dealing with hate speech. Ultimately, a CIB network advocating for hate speech could avoid detection by falling between the two departments (such a case of a CIB network pushing a pro-Iranian government narrative was brought before Meta’s Oversight Board in June 2026: oversightboard.com ↗).

The dream of countering Influence Operations by relying solely on behavioural and technical signals was already fantasmatic, as our first version of the CIB detection tree shows: some degree of content and narrative analysis is unavoidable.

But our new study found that platforms themselves are now lowering the barrier between content-narrative analysis and behaviour-anti-spam detection.

THE DISAPPEARANCE OF THE “CONTENT / BEHAVIOUR” SEPARATION

In 2017, Facebook was founding the CIB doctrine in these words: *“False amplification, which we define as coordinated activity by inauthentic accounts with the intent of manipulating political discussion [...]. We detect this activity by analyzing the inauthenticity of the account and its behaviors, and not the content the accounts are publishing.”* — Weedon, Nuland, Stamos, Information Operations and Facebook ([about.fb.com ↗](https://about.fb.com/)), 27 April 2017, p. 6

Almost 10 years later, the content / behaviour separation is eroding.

This is the most enlightening lesson from this update regarding the platform policies: the strict frontier between Content and Behaviour is less relevant for the platforms. In the policies, they tend to blur, and the CIB detection measures can no longer rely solely on behavioural indicators, while content moderation is incorporating automated techniques.

Meta was the company most insistent on separating Behaviour detection and Content analysis. But even Meta’s CIB measures now openly say they do look at narratives (even though they claim it is not taken into consideration).

Coordinated Inauthentic Behavior (CIB)

CIB remains a foundational pillar of Meta's adversarial threat reporting. CIB involves coordinated networks of assets working together to mislead people using false identities. Our enforcement is rooted in deceptive behavior and coordination, not the content posted—regardless of the narratives promoted or whether an operation is foreign or domestic. We also continue to monitor for reconstitution and remove assets connected to previously disrupted networks. Our reporting in this section contains details of CIB networks that we removed in the fourth quarter of 2025.

CIB topline this half include:

- **Two deep dives illustrating evolving tradecraft:**
 - **We investigated and removed an influence operation originating in Iran and targeting the US**, which used sophisticated fake personas on Instagram to misrepresent identity and amplify narratives. The case illustrates how influence operators increasingly blend amplification with attempted relationship-building.
 - **We disrupted a Pakistan-based domestic-targeting operation anchored in off-platform websites and coordinated amplification by pages and personas.** The case highlights the use of AI-assisted content generation across languages to increase output quality and consistency, and the use of identity misrepresentation (including stolen imagery) to appear locally authentic while promoting nationalist narratives and disparaging perceived adversaries.
- **Beyond the deep dives, we continued to investigate and remove other CIB networks, including operations originating from Russia, China, and Iran**, disrupting recurring patterns such as impersonation of grassroots media, localized branding to appear domestic, and cross-platform presence used to reinforce credibility.

Figure 1. Meta, Adversarial Threat Report, First Half 2026, section "Coordinated Inauthentic Behavior (CIB)".
transparency.meta.com ↗

To fight scams, Meta has to increasingly combine behavioural signals and content analysis. In its H2 2026 Scam section, Meta admits

"the signals that matter most are often behavioral rather than content-based. Dormant accounts produce no content to classify. Engagement bait looks indistinguishable from ordinary posts. Fake Charity pages solicit donations without ever posting a link that a URL scanner would flag. In each case, it was coordination patterns, structural anomalies, or partner intelligence—not content analysis alone—that exposed the operation"

Two main reasons explain this shift from a strict separation to a more hybrid approach.

1. The obligation to fight electoral manipulation

The threat of electoral manipulation on the platforms became a top priority during 2024, *“the big election year”*. During that time, higher pressure was put on the platforms to monitor electoral manipulation.

The Code of Conduct on Disinformation, integrated into the Digital Services Act under its Article 45 in February 2025, in its *“Common understanding of impermissible manipulative behaviour”* (inherited from the 2022 Strengthened Code of Practice) explicitly lists 8 TTPs characteristic of *“Impermissible Manipulative Behaviour”*. Among these 8, three are defined by content (hack-and-leak operations, malicious deepfakes and non-transparent paid messages or promotion by influencers), one by identity (impersonation) and four by behaviour.

The *“Commission Guidelines for providers of Very Large Online Platforms and Very Large Online Search Engines on the mitigation of systemic risks for electoral processes”*, issued in April 2024, specifically asked the platforms to establish multidisciplinary teams to counter electoral manipulation, explicitly asking the platforms to mix fact-checking, content moderation, disinformation and behaviour-based FIMI:

To reinforce internal processes and resources in a particular electoral context, providers of VLOPs and VLOSEs should consider setting up a dedicated, clearly identifiable internal team prior to each individual electoral period (...) The team should cover all relevant expertise including in areas such as content moderation, fact-checking, threat disruption, hybrid threats, cybersecurity, disinformation and FIMI, fundamental rights and public participation and cooperate with relevant external experts, for example with the European Digital Media Observatory (EDMO) hubs and independent fact-checking organisations

eur-lex.europa.eu ↗

Bluesky itself rewrote its community guidelines in August 2025 around four principles, one of them *“Be Authentic”*, an update it explicitly presents as aligning its harm categories with the UK Online Safety Act and the EU Digital Services Act: the restructuring and the regulatory alignment come in the same move bsky.social ↗

But this is not just theoretical. In practice, the focus on the elections forced the platforms to mix content analysis and behavioural signals. In the various threat reports published by the platforms, one can see that the chapters dedicated to this particular threat include both CIB detection and content analysis.

2. AI

AI increased the capabilities offered to both adversaries and the defenders. LLMs can now generate more original-looking content (if not authentic), and AI agents can be used to operate multiple accounts or networks of accounts independently, cross-platform. AI also offers new

capabilities to analyse the content and track narratives, rendering the frontier between narrative and behaviour methodology less and less relevant.

How this changes detection on both sides, and how the AI companies themselves detect, is developed in the last chapter, *"CIBs and AI"*.

ARE CIBS POLITICAL OR COMMERCIAL?

CIBs are mostly created for commercial reasons. The algorithm favours what has the most followers, likes or comments, so being backed by a small army of inauthentic accounts would be the preferred solution for anyone who wants to boost their content. But CIBs are also, for the very same reason, political. Anyone willing to boost political content would use an army of inauthentic accounts.

Political CIBs are easily detected and removed (for example the Zemmour campaign in France in 2022, whose team amplified twelve petitions on Twitter and Facebook through coordinated cross-posting, as documented by ISD: isdglobal.org ↗). Political involvement is highly sensitive; political content and election periods are under high scrutiny.

Conversely, commercial CIBs flourish and thrive under the radar. First, because they are believed to be of much less consequence, and also because if platforms had to really remove every one of the inauthentic accounts, their numbers would drastically shrink. So platforms have to play on the edge: too many inauthentic accounts allowed on a platform and advertisers would not trust you to sell ads to real people. Too many inauthentic accounts removed on a platform and the advertisers will pay less for a smaller reach. One exception to that rule is X, which drastically reduced its anti-CIB enforcement after its acquisition in October 2022 (half of its election integrity team was cut in September 2023 theregister.com ↗; X's global trust and safety staff had been cut by a third and its safety engineers by 80% since the acquisition, esafety.gov.au ↗; and a study of X from January 2022 to June 2023 found *"no reduction in inauthentic activity"* and *"a statistically significant increase in the bot scores of randomly sampled accounts after Musk's purchase"*, Hickey et al., PLOS One, 12 February 2025, doi.org ↗). Consequently, X saw its advertising revenue fall to about \$2.5 billion in 2023, against more than \$1 billion per quarter in 2022 ([Bloomberg](#), via [Fortune](#), 12 December 2023).

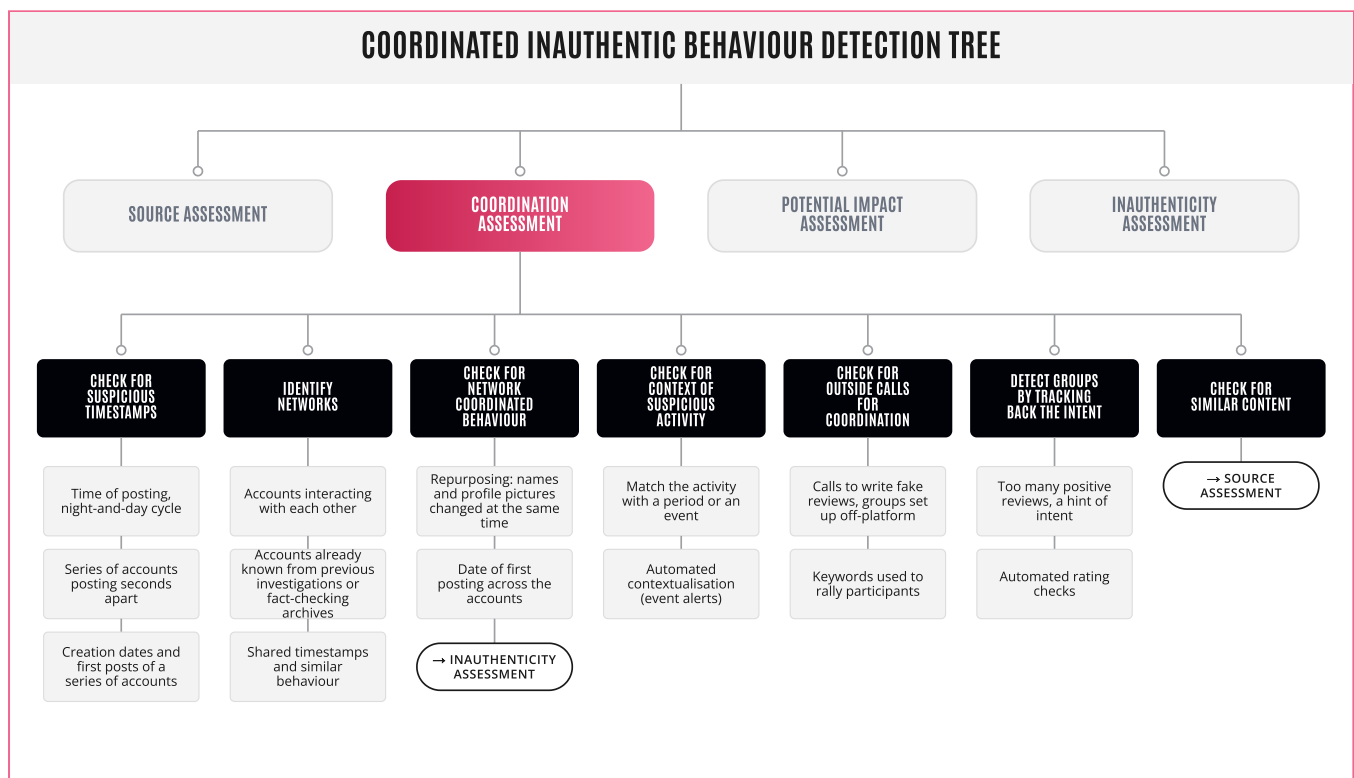
At the other end of the spectrum, review-based platforms follow the opposite model: Yelp, Tripadvisor or Amazon. Their whole business model is based on authentic reviews posted by authentic accounts, so their policies and detection methods are much more severe.

In reality, the commercial and political CIBs are deeply intertwined. Political actors hire commercial networks to help them and now we see the reverse phenomenon: commercial actors using political content for monetisation. Facebook was already identifying the trend in 2021 that they dubbed *"IO as a service"*: *"Commercial actors offer their services to run influence operations both domestically and internationally, providing deniability to their customers"* about.fb.com ↗ What was a threat in 2021 has now become a normal behaviour for CIBs and reaches the industrial level ladn.eu ↗

WHY A DETECTION TREE

This work is based not on the CIB technique but on the detection methodology. A list of all CIB techniques would be unreadable and would end up with 200 or more techniques, impossible to classify and difficult to bear in mind for analysts. Here, the detection tree is articulated around two dozen detection methods that are much more stable over time and with one method that can reveal 25 techniques, it becomes possible to quickly cover everything and even detect future techniques.

Taxonomies of different techniques used by CIB networks are always possible but most of the time useless. Techniques change, networks adapt, new techniques emerge and old ones disappear but never completely. Focusing on detection methods allows us to produce a readable methodology that can be used by analysts. A behaviour caught by several different detection methods will be tied to different branches of the methodology, giving it the “tree” aspect.



The Coordination assessment branch, as an example of how detection methods tie behaviours to the branches of the tree. Criteria from the 2021 publication; the pills point to the branch where a criterion continues.

DETECTION METHODOLOGY

The detection methodology relies on two pillars: Accumulation of Indicators and Attackers' Mistakes.

ACCUMULATION OF INDICATORS

There must be a network of accounts to generate coordination between them. One inauthentic behaviour does not reveal anything. Accounts can be real or fake; it is of no consequence for the methodology. The world of CIB is a great sea of nuances between *"real"* or *"fake"* accounts. One will find real accounts hijacked or hacked to serve a CIB campaign, real individuals paid to use their own accounts to participate in a campaign, a real person operating a whole network of accounts, or AI carefully designing fake accounts that look real. Focusing on Inauthenticity rather than fakeness spares the analysts and the investigators from having to prove the fakeness. These nuances can be categorised into 4 big families:

- Human behaving like a Human

- Human behaving like a Bot

- Bot behaving like a Human

- Bot behaving like a Bot

Human behaving like a Human will be the home of false positives. Every other category is potential CIB.

To track CIBs, one has to accumulate a series of indicators, shared by the same network, and answering 4 big questions:

- Who are the accounts? (source assessment)

- How do they operate together? (coordination assessment)

- What are their numbers? (impact assessment)

- Do they look fake? (inauthenticity assessment)

ATTACKERS' MISTAKES

The use of AI undeniably increases the possible quality of the CIB networks.

anthropic.com ➤ Anthropic reports that

“Threat actors find ways to obscure the origins of their identities. This began at the outset of the content creation process: actors asked the model to strip the marks of automated text and to sound organic, built account warmup and evasion logic, and removed metadata and codenames before delivery. They also laundered their access to Claude itself through VPNs, foreign phone numbers, rotated accounts, and third-party services that masked their IP address.”

However, the true purpose of CIB remains quantity, not quality. Attackers do not really care about their accounts and often need to mass-produce them rather than carefully protect every account by creating realistic biographies, profile pictures or differentiating every piece of content by renewing the prompts. AI offers attackers new ways to improve the quality of inauthenticity and concealment of their campaigns, but quality is not what attackers look for.

As a consequence, mistakes, obvious tells and clear giveaways are commonly found in CIB investigations. Reset’s *“monetized slopaganda”* found that the Gemini watermarks were not removed from the deepfake videos, or that prompts in Urdu were directly copy-pasted into the content of the posts.

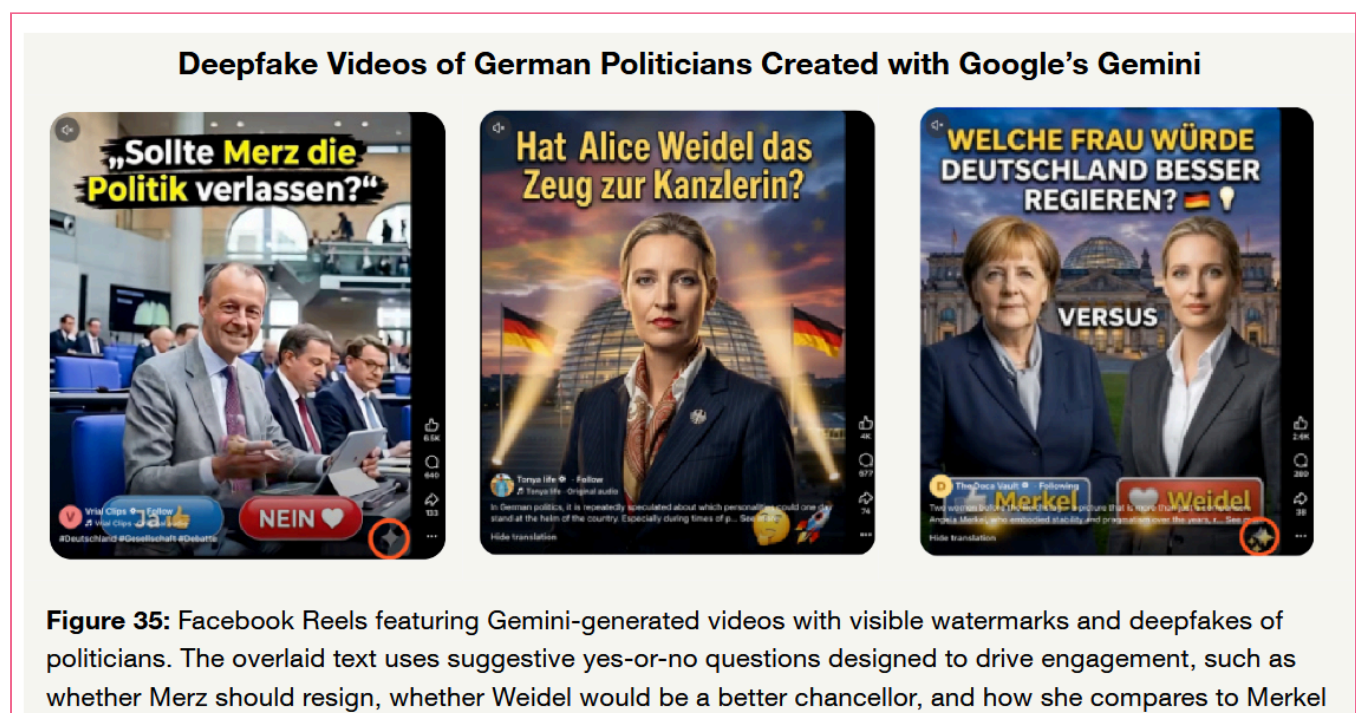


Figure 2. Reset Tech, Monetized Slopaganda, September 2026, Figure 35. reset.tech

Urdu is the Language of the Operators, as Exposed in ChatGPT Prompts



Figure 44: Copy-and-paste errors containing prompts and follow-up suggestions in Urdu (highlighted in orange) and preambles in German (highlighted in green). The first post begins with “OK, Irfan” and contains a German-language preamble, followed by a suggestion in Urdu: “If you want, I can make it more emotional or in a more journalistic tone.” The second post contains prompt instructions in Urdu beginning with, “Brother, for this image, create text according to what is shown in the picture,” followed by instructions to produce “engaging,” “not boring,” and “a full paragraph of 13 lines” (Source: Facebook 1, 2).

Figure 3. Reset Tech, Monetized Slopaganda, September 2026, Figure 44. reset.tech ✓

The French VIGINUM agency found out that one campaign against a political candidate misspelled his name.

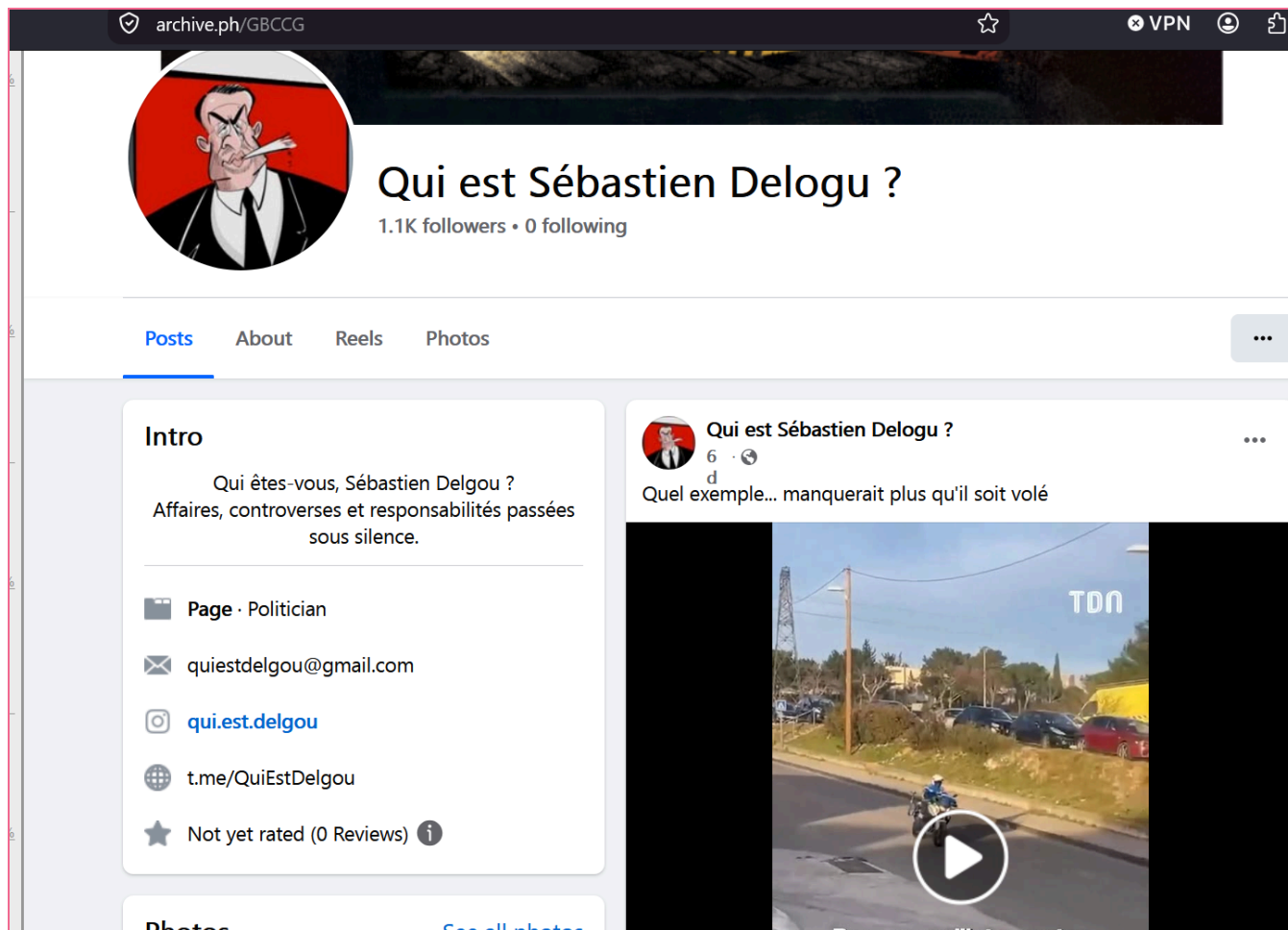


Figure 4. Facebook page "Qui est Sébastien Delogu ?", archive cited by VIGINUM (archive.ph ↗) in its Rokh Solis technical report, June 2026, misspelling "delgou" in the email address and in the link to the Telegram channel. sgdsn.gouv.fr ↗

DEFINITIONS OF CIB BY THE PLATFORMS

Definitions of what constitutes CIB for the platforms vary greatly. Studying the definitions, however, can help identify what the target of the detection methods is. The most important difference between platforms is between those allowing pseudonymity and those demanding that a real name and real identity be used. This difference is then divided between platforms tolerating multiple accounts and those demanding one account per person.

Definitions of what is allowed and what is not, thus defining what is CIB and how platforms track it, can be found in the Terms of Service and the Community Guidelines. Those policies cover everything that is banned and do not look at the content produced, thus tracing a clear line between content and behaviour.

Finally, a big part of the anti-CIB measures from the platforms recycles or updates their automated systems against spam. Therefore, a big part of what platforms consider CIB and how they detect and track them can be found in their anti-spam policies.

FACEBOOK

Facebook has two main policies:

In the Terms of Service, Facebook demands that real identity and real information be used to create an account, and only one account per person.

Facebook also has in their community standards an *“Authentic Identity Representation”* policy specifying that they will consider it suspect to:

Provide a false date of birth

Use a name that is not the one used in everyday life

Multiple connections to the same account

Accounts representing a non-human entity (fictional character, pet or business)

Prolonged Dormancy

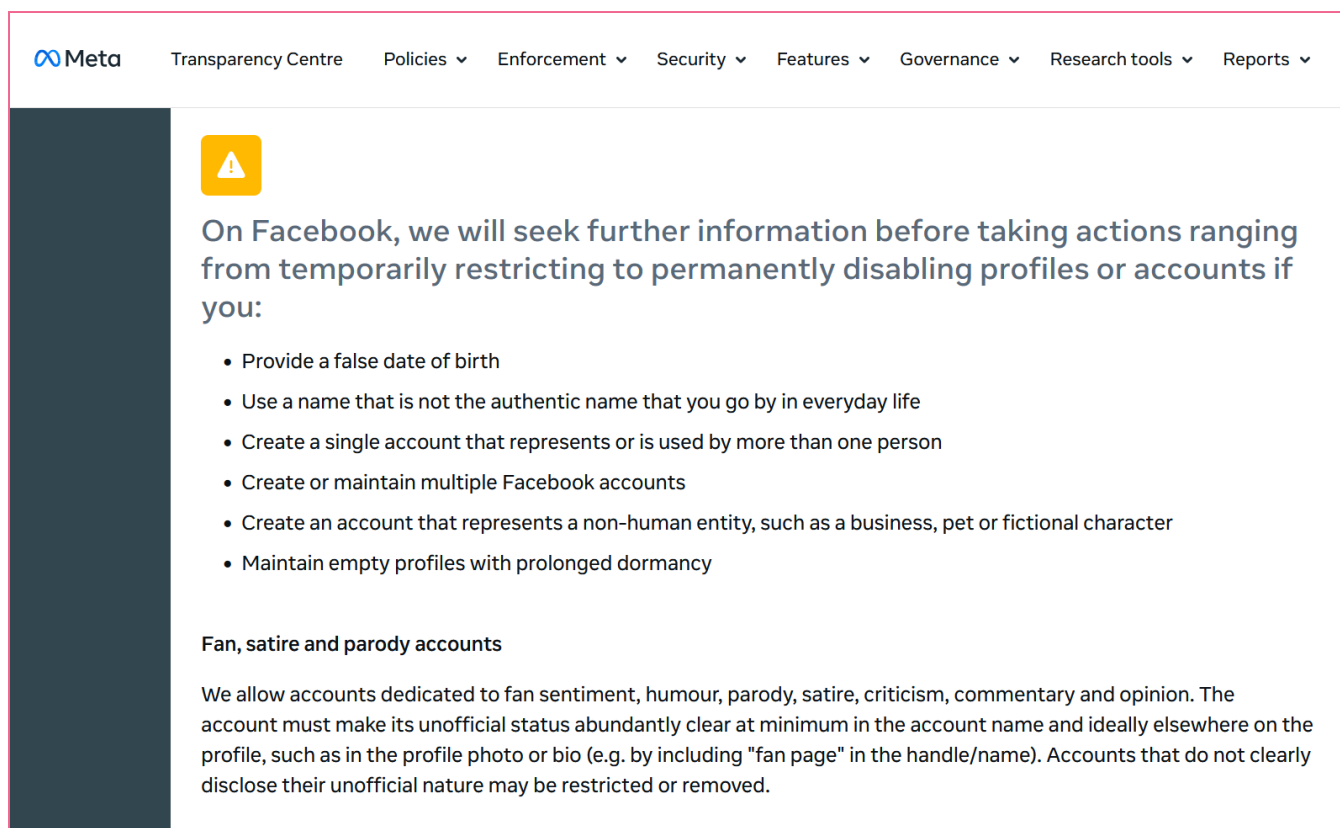


Figure 5. Meta Transparency Centre, Community Standards, "Account Integrity and Authentic Identity".
transparency.meta.com ↗

What they track

Repeated or significant changes to identity details, such as name or age

Misleading profile information, such as bio details and profile location

Using stock imagery

Through its anti-spam policy, Meta tracks a high frequency of any action: posting, sharing, creating accounts, groups, events.

Meta also tracks low frequency when associated with other indicators.

They also track content, if they detect attempts to buy or sell accounts, pages, engagement (likes, views etc.) or content requiring people to like or share before being able to win something or see some content. It also tracks changes in admin or moderator roles in groups or pages.

Meta tracks URLs and links on its platforms and outside domains. It checks if the URL and the content outside match the content delivered on the platform. They check if the URL changes, redirects or requires an action when it lands (captcha, watch ad, click here).

Furthermore, Meta measures inauthentic metrics as they do not allow *"content that is designed to deceive, mislead, or overwhelm users in order to artificially increase viewership."*

transparency.meta.com ↗

Finally, Meta tracks the attempts to *“evade enforcement under the Community Standards”*; it is even central to its definition of CIB. Therefore, it tracks the identity of the account creator to check if it is linked to previous violations or if it attempts to circumvent restriction measures.

Meta also tracks coordination and will remove or take action against accounts that can be linked to, associated with or have *“close linkage with a network of accounts or other entities that violate or evade our policies”* transparency.meta.com ↗

Facebook also bans scraping, and through that policy, checks what it considers *“automated means”* and *“regardless of whether such automated access or collection is undertaken while logged-in to a Facebook account”*, which means tracking anonymous traffic and connections without being logged in.

Facebook also tries to put barriers at account creation: *“A simple example is blocking certain IP addresses altogether”* about.fb.com ↗

Finally, Facebook also relies on accumulating different signals as a way to prevent false positives. They call it *“Deep Entity Classification”*, and use AI to detect inauthenticity and coordination signals transparency.meta.com ↗

INSTAGRAM

Although Instagram belongs to Meta, you are allowed to provide the identity you want. *“You don't have to disclose your identity on Instagram, but you must provide us with accurate and up to date information (including registration information), which may include providing personal data.”*

Instagram is also wary of bots created for scraping.

“You can't attempt to create accounts or access or collect information in unauthorized ways. This includes creating accounts or accessing or collecting information in an automated way without our express permission”.

help.instagram.com ↗

It is forbidden to sell or purchase Instagram accounts, implying Instagram checks for username transfer and credential sharing.

Instagram does not like metrics manipulation. Thus, it tracks the activity of third-party apps connected to the accounts and will enforce measures against accounts using apps to increase their metrics help.instagram.com ↗

However, Instagram restrains itself from enforcing these policies too harshly as they cannot fully prove or attribute this behaviour to an intentional violation of their policies. Inauthentic accounts do follow and like random people to appear authentic and Instagram cannot directly prove it

comes from the account owner: “Sometimes, non-Instagram apps create automated likes, comments and follows to make an account appear more popular than it is.” help.instagram.com ↗

In March 2023, the then French Prime Minister was trying to force a law through parliament using a controversial provision named 49.3. To mock her, an Instagram user paid a third-party service to increase the Prime Minister’s Instagram account followers so it reached 49.3k followers radiofrance.fr ↗

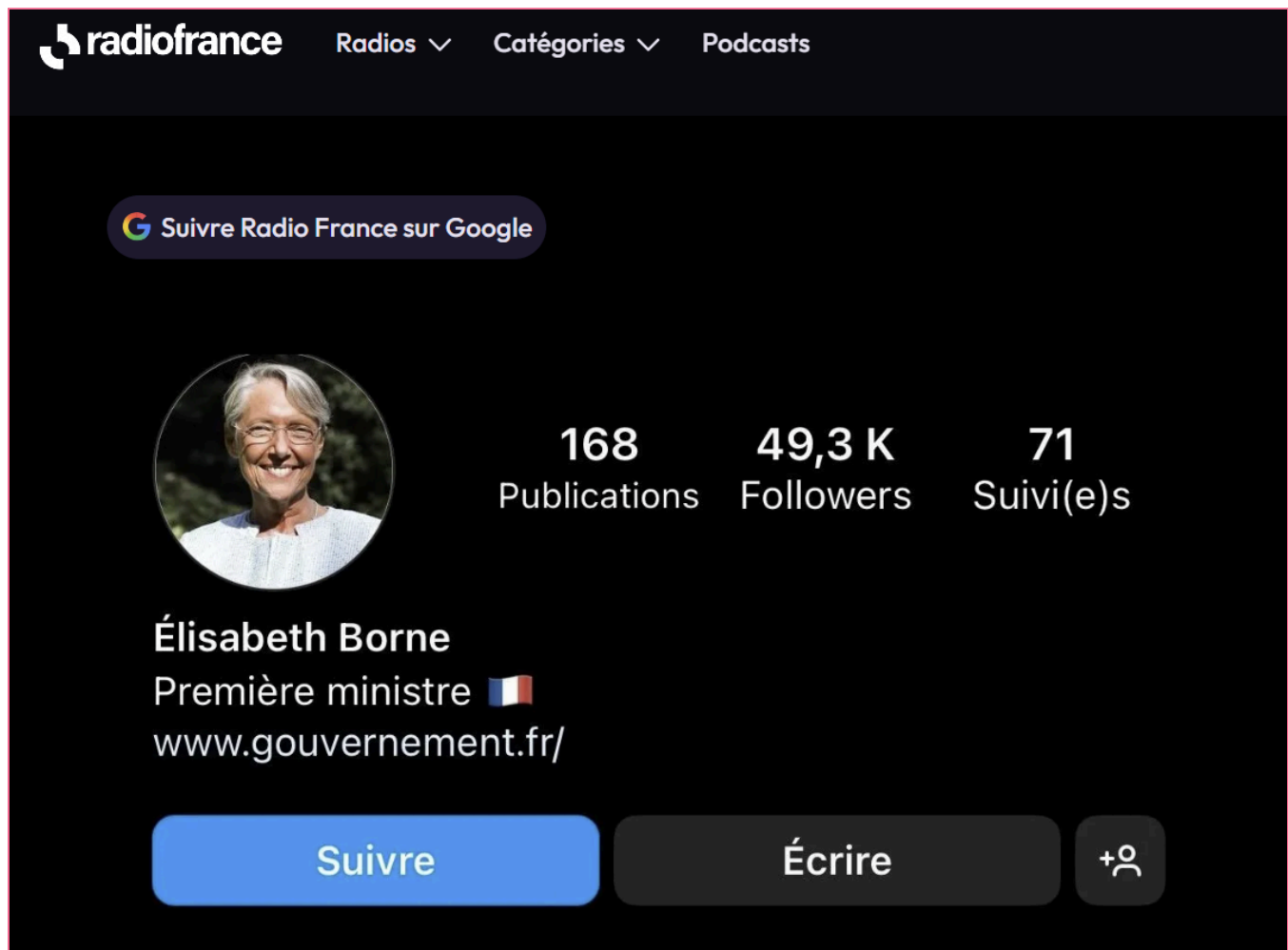


Figure 6. The then Prime Minister's Instagram profile showing “49,3 K” followers, as reproduced by France Inter, 20 March 2023. radiofrance.fr ↗

Instagram’s Spam policy redirects to Meta’s community guidelines.

X (FORMERLY TWITTER)

For X, “accounts must be authentic”. In practice, it means accounts should not use “unauthorised automation”. X tracks mass registration and the origin of the accounts to enforce this policy.

X bans “manufactured identities to engage in disruptive or deceptive behavior. This may include using stock, stolen or AI-generated profile photos, copied or stolen profile bios, and/or misleading

profile information for the purpose of deceiving others." It would tend to imply that X checks the profile picture provenance (through AI detection) and bio text to check if they are copy-pasted.

Multiple accounts are tolerated but X prohibits *"coordinated inauthentic activity"* and checks if the accounts operate *"with the same or substantially similar content"*, in order to manipulate metrics or inflate engagement. X thus checks if multiple accounts attempt to engage with the same posts or create self-amplification loops by boosting one another. Up to 10 accounts per phone number are allowed (help.x.com ↗). X also allows *"Authentically coordinating with others to express ideas, viewpoints, support, or opposition towards a cause in a non-violative manner."* This means coordination alone is not a criterion.

Attempts to circumvent or *"evade X enforcement actions"* are also monitored, and this policy gives many indications of what the platform is tracking:

- Creating new accounts
- Imitating a suspended account to replace it
- Repurposing an already-existing account
- Having someone else operate an account on your behalf

The *"Inauthentic Behaviour"* section is mostly about spam behaviours: Sending bulk, aggressive, high-volume unsolicited mentions or direct messages, using popular hashtags to divert traffic, repeatedly posting links without commentary, posting and deleting the same content repeatedly, or copy-pasting the same or nearly the same message.

A new and interesting prohibited behaviour appeared with the authorisation to edit posts: *"deceptively editing"* (to amplify content that differs from the original post) and editing links so the final destination page changes.

In what X calls *"engagement spam"* one can find the list of prohibited coordination.

Coordination to exchange engagement (likes, polls, replies etc.). Coordinating or compensating others to like, share, etc.

Engaging in what X calls *"Follow churn"*: following and unfollowing a large number of pages

Following too large a number of people in a short time or automated following

Duplicating another account's followers (with automation)

Submitting false reports in large numbers (unless you are a law enforcement organisation)

X also examines content behaviour to track *"Inauthentic Content"*: mostly content that has been *"substantially edited or post-processed in a manner that fundamentally alters (...) and distorts their meaning"*, or the use of algorithms and AI to fabricate a real person.

In that category also falls media content presented with false or misleading context (source, time, location, misquotation etc.).

The *“How We Enforce”* section lists the enforcement measures X might take if they detect a violation:

Anti-Spam challenge, Denylisting URLs, restricting reach, Temporary loss of access to X features or products, Profile modification (they may require you to edit the content of the profile), and suspension. Nearly all of these measures rely on automated means apart from profile modification and maybe suspension, meaning CIB detection for X heavily relies on automated systems.

As X is also the base platform for Grok, Elon Musk’s AI, the rules concerning automation are more detailed on this platform. Although Grok itself is a product of SpaceXAI and the usage of Grok on X is supposed to follow the X policies ([x.ai ↗](#)), X policies had to allow some degree of automation on its platform. *“Provided you comply with all other rules, you may leverage artificial intelligence (AI) technologies to create automated reply bots that generate dynamic, context-aware responses, as these can enhance user engagement, provide timely assistance, and foster innovative interactions on X.”* The overall policies on automation ask users to comply with the rules and to use automation to do good, like *“Try new things that help people (and comply with our rules)”* [help.x.com ↗](#) and not to do bad, like *“violate these or other policies(...)”*.

Specifically for Grok, it is forbidden to use Grok to create *“fake accounts”*, to be *“Misleading others or not being transparent regarding your use of AI, including by phishing, creating fake accounts, providing services that appear to be from you, when they are in fact from SpaceXAI, or providing services that appear to originate from SpaceXAI, when they do not”*. These policies are listed under the *“comply with the law”* section.

X uses automated means to detect content violating its moderation rules at a very basic level: it simply instructs LLMs to detect certain keywords:

“X employs a combination of heuristics and machine learning algorithms to automatically detect content that we believe violates the [X Rules and policies](<https://help.x.com/en/rules-and-policies/x-rules>) enforced on our platform (...) Heuristics are common patterns of behaviours, text, or keywords that may be typical of a certain category of violations. Pieces of content detected by heuristics may also get reviewed by human content reviewers before an action is taken on the content.”

This is, according to X, assisted by teams of human investigators for identifying Tactics, Techniques and Procedures [transparency.x.com ↗](#)

BLUESKY

All of Bluesky's CIB policies can be found in the *"Community Guidelines"*. But one automation technique found in the ToS is of interest to CIB detection: Bluesky partly automated the detection of illegal content through recognising hashes of known illegal content: *"This proactive technology includes tools that help us recognize hashes of known illegal content and that can detect violations of our terms based on the text in and content of a post"* [bsky.social](#) ✓

"Be Authentic" is one of the 4 core principles of the community guidelines.

Bluesky also monitors its features to check that they are not abused for automated harassment: *"Do not abuse Bluesky features (lists, labels, community moderation tools), engage in bad-faith mass reporting, use automated harassment systems (...)"*

The very first prohibited behaviour listed in the *"Be Authentic"* section is *"do not send spam"*, indicating that here too, the anti-spam policies are at the heart of the anti-CIB policies. Multiple accounts are allowed: *"using alternative identities without misleading other users"*. It is prohibited to *"create fake accounts to deceive others about who you are"*, and to *"engage in deceptive account practices"* like *"identity churning (changing your account identity to keep followers)"* and *"coordinated deception"*.

It is prohibited to *"abuse platform trust & safety systems by: falsifying documents or misrepresenting your verification status"*, meaning Bluesky relies mostly on entry barriers to protect authenticity on the platform.

Trust and Safety reports are indeed where the anti-CIB fight occurs [bsky.social](#) ✓ and also where new automated solutions are developed and tested. *"We currently register users for additional defenses when we see a pattern of new account harassment, but in the future, we'll be able to better detect and surface when multiple new malicious accounts are created and managed by the same user."* Content is also examined and Bluesky develops a system to label *"rude"* content. Bluesky also uses this automated system on other content categories: *"In December [2024], we were able to review our first wave of automated reports for content categories like impersonation."* [bsky.social](#) ✓

On anti-spam, Bluesky is *"launching a pilot project to automatically detect when an account is clearly fake, scamming, or spamming users to hopefully reduce the likelihood this happens."*

In age verification and in complying with legal requests, Bluesky opted for a country-by-country solution. This means Bluesky has to seriously assess the geolocation of its users.

Although Bluesky relies heavily on automated systems, the platform is innovative in mixing anti-CIB with harmful content detection and moderation, leading to the development of automated solutions to analyse content. They differentiate content and behaviour not in the detection but in the enforcement of measures. *"When violations involve specific content, we remove the problematic material while preserving the account. When violations involve bad-faith actors —*

impersonators, spam networks, coordinated manipulation — account-level action becomes necessary” [bsky.social](#) ↗

“Our automated detection systems identify potential violations before users encounter them or report them. These systems focus on categories where we can either identify patterns in behaviour (like spam patterns and known bot attack signatures) or where content is harmful (such as Child Sexual Abuse Material via CSAM hashes)” [bsky.social](#) ↗

The same automation philosophy is applied to *“Influence Operations”*, where Bluesky also works on identifying coordination: *“Influence operations are coordinated campaigns that attempt to manipulate public discourse through deceptive or inauthentic behavior”* [bsky.social](#) ↗

In this work, Bluesky admits to working with *“public-sector partners and independent researchers to investigate and mitigate suspected influence operation campaigns”*. Bluesky uses automated systems to identify the networks based on *“behavioral patterns and technical indicators—such as coordination signals and account relationships”*, then investigates with its human teams or external partners.

What they were able to detect was:

impersonation of journalists and researchers,

use of inauthentic or misleading account identities

leveraging interoperability (and using third-party services to operate accounts from other platforms)

spam manipulation with Follows and Starter Packs for visibility and reach.

LINKEDIN

LinkedIn sets the bar higher in terms of real identity and fake accounts. The user agreement itself specifies that *“you agree that you will (...) Provide accurate contact and identity information to us and keep it updated; Use your real name on your profile”*. It strictly forbids *“Create a false identity on LinkedIn, misrepresent your identity, create a Member profile for anyone other than yourself (a real person), or use or attempt to use another’s account (such as sharing log-in credentials or copying cookies)”*

Robots are strictly forbidden for scraping but also for creating *“inauthentic engagement”* on the platform. *“You agree that you will not (...) Use bots or other unauthorized automated methods to access the Services, add or download contacts, send or redirect messages, create, comment on, like, share, or re-share posts, or otherwise drive inauthentic engagement;”*

LinkedIn policies are divided into three pillars: Be Safe; Be Professional; Be Trustworthy.

In the “Be Safe” pillar, one will find policies on content and moderation.

Be Trustworthy addresses CIBs: *“We require you to use your true identity on LinkedIn, provide accurate information about yourself or your organization, and only share information that is real and authentic.”* It is here again specified that LinkedIn does not accept *“fake profiles or entities”*. It is prohibited to use a picture not depicting oneself, to lie about professional affiliation and to use someone else’s account or to share an account with anyone else. This implies LinkedIn checks profile pictures and multiple connections.

Content and anti-CIB are here mixed. Concerning the content, it is forbidden to *“share content to interfere with or improperly influence an election or other civic process.”* AI content or deepfakes are forbidden unless specifically labelled. And it is forbidden to share *“content that directly contradicts guidance from leading global health organizations and public health authorities.”*

This means LinkedIn is specifically monitoring health and electoral propaganda content.

The “spam” section of “LinkedIn Help” tells a bit more about the means designed against CIBs. [linkedin.com](https://www.linkedin.com/help/linkedin/answer/51888) ✎ *“We may remove or limit the distribution of content designed to artificially increase engagement through misuse or misrepresentation of LinkedIn’s features.”* In Example of Spam, they list:

“Emoji / reaction polls that artificially boost engagement; Widely circulated “chain letter”-type content, requesting likes, reactions, and shares; Excessive, irrelevant, or repetitive comments or messages; Articles that are misleading or malicious, often in order to manipulate search engines.”

This last example is explicitly fighting an activity designed to manipulate other platforms, which is necessary against CIBs, as many behaviours are designed to be cross-platform (see the Coordination branch of the detection tree).

LinkedIn transparency reports give numbers on how many accounts they “stopped” but without providing much about the techniques they use. [about.linkedin.com](https://about.linkedin.com/en/transparency) ✎

Nearly 80% of the “fake accounts” are “stopped at registration”, meaning LinkedIn relies heavily on anti-spam as its main line of defence.

LinkedIn’s defence against “fake accounts” goes back to 2018 and they built from there.

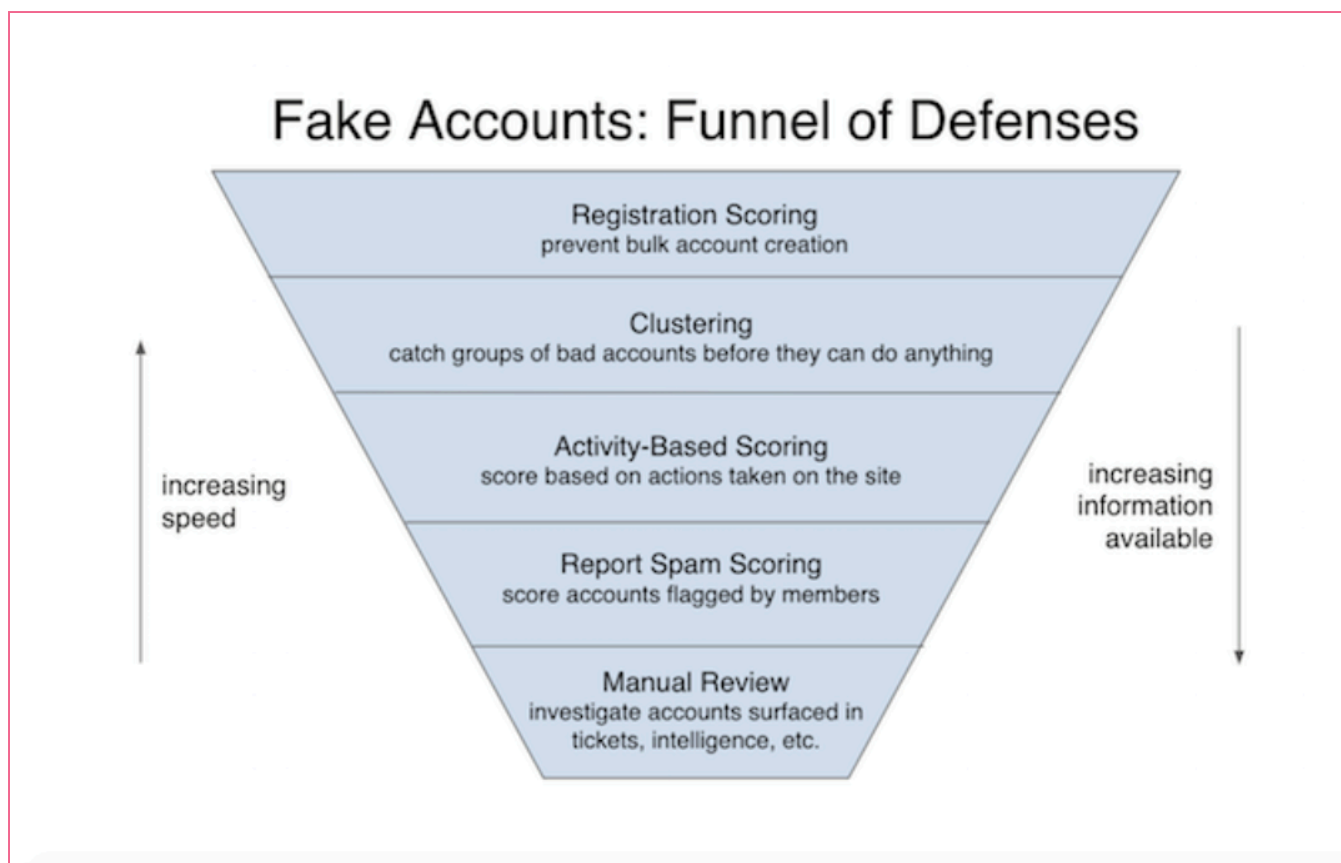


Figure 7. LinkedIn Engineering, "Automated Fake Account Detection at LinkedIn", 12 September 2018.
[linkedin.com](https://www.linkedin.com/company/linkedin-engineering) ↗

They attribute a score at registration and they also do "*clustering*", trying to group together accounts that display similar behaviours.

Today, LinkedIn is engaged in fighting AI slop, which changes the way platforms address CIB policies: they need to look at the content with automation tools and AI detectors. LinkedIn uses "*technology systems built in partnership with our editorial team that have been trained to recognize signals of AI slop and learn over time by identifying content that adds perspective, context, or expertise and content that feels generic or repetitive*" news.linkedin.com ↗

LinkedIn also tracks coordination through what they call "*engagement pods*", groups of users coordinating, often off-platform, to boost each other's content. They check the frequency and rate of comments: "*if we detect excessive comment creation or use of an automation tool, we may limit the visibility of those comments*". socialmediatoday.com ↗ This is relatively new and was added in late 2025.

Later, LinkedIn decided to track third parties or tools used to comment or manipulate engagement on LinkedIn: "*Automated comments - comments submitted to LinkedIn via a browser extension, script, or third party tool, without human action involved in clicking the "comment" button - are not allowed on LinkedIn.*" [linkedin.com](https://www.linkedin.com) ↗

Countermeasures from LinkedIn are primarily to reduce the reach and visibility of this type of content.

WHATSAPP

WhatsApp prohibits *“sending illegal or impermissible communications such as bulk messaging, auto-messaging, auto-dialing, and the like”*. This means WhatsApp probably checks the frequency and volume of messaging.

The use of automated means is widely prohibited, especially to *“create accounts for our Services through unauthorized or automated means”*, and of course scraping (*“collect information of or about our users in any impermissible or unauthorized manner”*) and it is specifically forbidden to share an account: *“distribute or make our Services available over a network where they could be used by multiple devices at the same time”*, meaning WhatsApp checks for multiple connections.

“We remove over two million accounts per month for bulk or automated behavior — over 75% without a recent user report. These efforts are particularly important during elections where certain groups may attempt to send messages at scale” (WhatsApp, *“Stopping abuse: How WhatsApp fights bulk messaging and automated behavior”*, white paper, 2019, linked from faq.whatsapp.com ↗)

WhatsApp is also interesting as they are forced to ignore content and to focus on behaviour: *“Given the nature of end-to-end encryption, we focus on the behavior of accounts and user-reported content”*.

The first thing they detect is the location of the phone number and the origin of the SIM card. WhatsApp can be more lenient toward phone numbers coming from places where they are difficult to acquire in bulk and focus more on *“Phone numbers originating from areas with a history of fraud may indicate to WhatsApp there is a problem with an account”* (WhatsApp, *“Stopping abuse: How WhatsApp fights bulk messaging and automated behavior”*, white paper, 2019, linked from faq.whatsapp.com ↗)

WhatsApp also focuses on automated messaging, as they know *“Sending mass messages from a mobile phone is time consuming”*. It is also easier to detect automation signals, as the abuser will have to build around the platform architecture and thus give away signatures: *“Attempting to distribute content en masse on WhatsApp requires users to work around the design of our platform. Doing so exhibits signals that we use to identify abusive accounts.”* The ban system seems completely automated but relies on gathering multiple signals, not just one: *“When we are confident that an account is abusive, we ban the account from WhatsApp altogether (...) identifying these accounts manually is not realistic. Instead, we have advanced machine learning systems that take action to ban Accounts.”*

Behaviour is studied at 3 levels: *“This abuse detection operates at three stages of an account’s lifestyle: at registration; during messaging; and in response to negative feedback, which we receive*

in the form of user reports and blocks". Whenever a user receives a message from a phone number not included in their contact list, WhatsApp asks the user if they accept or block the message. WhatsApp can then compile the blocking requests and identify a behaviour. "if an account accumulates negative feedback, such as when other users submit reports or block the account, our systems evaluate the account and take appropriate action".

WhatsApp also relies on blocking CIB at registration as its first line of defence but in a more subtle way:

WhatsApp also checks behaviour at account creation: proving the user owns a phone via a one-time code. But WhatsApp also specifies it provides *"optional two-step verification"*. This seems to indicate WhatsApp checks if a user uses 2FA and might list it as a good behaviour or trust signal.

They also check the IP address at registration to check if the IP is associated with already known suspicious behaviour.

The second line of defence is analysing the account behaviour. *"WhatsApp cannot see the content of messages passing through our system, though we are able to analyze the frequency of account activity"*. They established a baseline of normal behaviour for a user and compare to that:

"Normal users operate relatively slowly on WhatsApp, tapping messages one at a time or occasionally forwarding content (...) an account that registered five minutes before attempting to send 100 messages in 15 seconds is almost certain to be engaged in abuse, as is an account that attempts to quickly create dozens of groups or add thousands of users to a series of existing groups"

. This means WhatsApp checks the time of posting, the delay between posts, group creation, contact activity.

As for everyone else, CIB detection is a fine art and WhatsApp has to navigate false positives: *"In less-obvious situations, a new account might message dozens of recipients who do not have the sender's account in their contacts. This could be the beginning of a spam attack, or it could be an innocent user simply telling their contacts about a new phone number"*. Thus, WhatsApp has to rely on multiple signals, but still automated: *"We have advanced our technology to the point where we combine signals to make rapid determinations without human intervention"*.

Although WhatsApp can't access the content, they do address the intent of the CIB based on the behaviour: a suspicious link is more likely to emanate from an economically motivated CIB, while a disinformation or politically motivated CIB will *"seek to organize and drive interest within a particular country, region, or even city - often within a limited time window"* and attempt to message users without their consent.

Finally, WhatsApp tries to prevent abusive reporting. To do so, they look at the behaviour history between the reporter and the reported. Blocking or reporting an unknown user after one unsolicited message is normal behaviour; trying to report a user after several interactions is more

suspicious. When a user blocks another user without prior interaction, WhatsApp looks at whether the reporters had interactions between them to check for coordination.

WhatsApp also works cross-platform in two ways:

restricting the listing of group invite links by search engines (in 2020 WhatsApp asked Google to stop indexing invite links and added a *“noindex”* tag to its link pages, as reported by TechRadar: [techradar.com](https://www.techradar.com/news/social-media/whatsapp-noindex-tag) ↗)

checking content to see if someone claims to be able to evade detection or enforcement of measures on WhatsApp: *“off-platform information includes public claims from companies about their ability to use WhatsApp in ways that violate our Terms.”*

YOUTUBE

YouTube’s community guidelines are organised into 6 pillars, each of them subdivided into policies.

Spam, CIBs and behaviour are found in *“Spam and deceptive practice”*, while the five other pillars directly concern the content: *“Sensitive content”* (nudity, child safety, suicide, self-harm and eating disorders, and *“vulgar language”*), *“Violent or dangerous content”*, *“Regulated goods”*, *“Misinformation”* and *“Educational, Documentary, Scientific, and Artistic (EDSA) content”*.

The Spam and deceptive practice pillar is where most of the CIB policies of YouTube can be found. They are: Spam policy, Impersonation policy, External links policy, Fake engagement policy, Playlist policy and Additional policy. CIBs are mostly covered by the Fake engagement policy, Impersonation policy and Spam policy.

In all its policies, YouTube associates checking content and behaviour, and its primary concern is metrics manipulation: *“boost content engagement”*, *“inorganic promotion”*, *“artificially inflate live traffic (‘viewbotting’)*” or *“content featuring a creator purchasing their views from a third party”*.

The Spam Policy, where most of the behaviour monitoring happens, states:

“We do not allow content, metadata, or behavior designed to take advantage of the YouTube community, including spamming users or manipulating the platform in ways that mislead them. We also prohibit misleading viewers to boost content engagement or to flood community spaces with inorganic promotion. (...) This policy applies to all types of content on YouTube, including unlisted and private content, comments, links, posts and thumbnails, and coordinated networks of channels.”

Examples given by YouTube include a 10-second video promising a *“free unreleased movie”* that directs viewers to a URL in the description, and technical manipulation (such as speeding up audio, heavy filters, cropping) to avoid YouTube detection.

This means YouTube probably strives to train its automated systems to detect these specific techniques. YouTube is also fighting to counter new techniques allowed by AI: *“Using automated tools or AI to churn out high volumes of similar content with minimal changes”* like *“Channels that use the exact same background music and repetitive AI generated imagery across many videos, with each video reading out an AI-generated script”*.

YouTube also checks if the title or thumbnails match the content of the video to prevent *“Using maliciously misleading titles, thumbnails, descriptions, or imagery to trick users into clicking on a video that does not deliver what was promised.”*

In its examples, YouTube gives away what they check: Repetitive or templated content; artificially inflating engagement through bots, coercion, or offering rewards in exchange for likes, views, or subscribers; coordinated *“sub-for-sub schemes”* (meaning YouTube checks reciprocal liking, similar followers of channels, timestamps of metrics etc.); Posting high volumes of low-effort, repetitive polls (e.g., *“Click the heart if you like food”*) multiple times a day; Posting identical or similar *“check out my channel”* comments.

The core of the fake engagement policy is the user *“intent”*: *“We consider engagement to be legitimate when a human user's primary intent is to authentically interact with the content. We consider engagement illegitimate, for example, when it results from coercion or deception, or when the sole purpose of the engagement is financial gain.”*

Assessing the intent of a user can only really be done by checking the content.

YouTube is especially attentive to metrics manipulation, fake likes, third parties inflating views, purchasing metrics, etc. *“Don't post content on YouTube if it fits any of the descriptions noted below. Links to or promotes third-party services that artificially inflate metrics like views, likes, and subscribers”*. It is allowed to *“encourage viewers to subscribe, hit the like button, share, or leave a comment”* but forbidden *“Offering to subscribe to another creator's channel only if they subscribe to your channel (“sub4sub”)”*.

The impersonation policy prohibits *“unauthorized impersonation of a person, entity, or channel that may mislead viewers”*. This concerns *“deceptively copying the branding, content, or usernames of individuals, channels or entities”*, and also includes the use of AI to copy the voice or the likeness of an individual.

YouTube prohibits (and therefore checks) *“Using workarounds like typos, special fonts, or extra punctuation to trick viewers”*. They also check copying and similarities: *“The channel does not have to be 100% identical, as long as the intent is clear to copy the other channel.”* The examples address *“putting “real” or “official” into the channel name”*; Pretending to be a backup channel by adding words like *“Backup,” “Shorts,” or “Live”*; *“Using AI to make it look like a famous person is willfully participating in your video”*; *“Misrepresenting an individual by using an AI version of their voice and false metadata”*; or *“the use of deceptive misspellings”*.

These examples imply YouTube probably checks name similarities (visual, phonetic, misspelling, special characters), profile picture and banner similarity, string similarity of names, regular checks against established entities and big channels.

Other techniques can be found in YouTube's Terms of Service, especially concerning their detection of automation. *"If your channel has been restricted due to a strike, you must not use another channel to circumvent these restrictions. Violation of this prohibition is a material breach of this Agreement and Google reserves the right to terminate your Google account."* This means YouTube will check all channels associated with an account in case a channel gets a strike. This probably also involves giving a reputation score to Google accounts.

Obviously, YouTube checks if someone tries to access its services via automated means or bots or scrapers (unless they follow the appropriate channel).

YouTube policies seem almost entirely focused on manipulated metrics, which could imply that manipulated metrics are at the core of their anti-CIB policies. Detection goes as far as checking the content, but YouTube probably goes further and checks whether it can detect an intent of metrics manipulation in order to enforce its policies.

81,773 YouTube channels across 57 separate actions were involved in Coordinated Influence Operation Campaigns. disinfocode.eu ↗

TIKTOK

Because of its Chinese origins, TikTok has always been a prime suspect for being the vector of Influence Operations or Foreign Interference, to the point that the US Court of Appeals for the D.C. Circuit, upholding the 2024 law forcing ByteDance to *"divest or ban"* TikTok in the US, found compelling the government's justifications of *"countering China's data collection and covert content manipulation efforts"*, a formulation the Supreme Court repeated in January 2025 while confirming the law on the data-collection ground alone supremecourt.gov ↗

TikTok has always tried to prove its good intentions through clear and wide policies and ToS (whether these policies are enforced will not be the subject here).

TikTok's anti-CIB policies are spread between *"Integrity and Authenticity"*, the Terms of Service, the content policy, and a dedicated page named *"Countering deceptive behavior"* in its *"Policies and engagement"* section.

The *"Integrity and Authenticity"* page addresses most of the issues, and content analysis is a big part of it. For each issue, policies are divided into what is strictly forbidden, what is *"Not Allowed"* and would suffer a ban, and what is *"FYF ineligible"*, meaning content will stay online but without visibility.

The first issue addressed is *“misinformation”*. It includes *“hoaxes, misleading AIGC [AI-Generated Content], harmful conspiracy theories, and other false information related to public safety, crises, or major civic events”* and it is analysed by fact-checkers and experts to *“assess the accuracy of content”*.

TikTok policies go as far as defining what a conspiracy theory is (*“Beliefs about unexplained events, or claims that involve rejecting generally accepted explanations for events. This includes suggesting they were carried out by covert or powerful individuals or groups.”*) and the difference between health misinformation that could cause significant harm, *“discouraging professional care for life-threatening conditions (e.g., vaccine effectiveness)”*, and the ones that would cause moderate harm, *“such as false claims about treating non-life threatening conditions like the common cold”*, and allowing *“Conversations about climate policy, weather, or technology, as long as they don't deny or misrepresent scientific consensus.”* [tiktok.com](https://www.tiktok.com) ↗

Covert influence operations, impersonation, spam, fake reviews, and sharing hacked materials are prohibited together in a dedicated section of the Integrity and Authenticity policy.

Elections are a special focus with a dedicated team *“staffed by multi-disciplinary experts in democracy, elections, civil society, and technology who bring a wealth of policy, technical, and local expertise to our preparedness efforts.”* TikTok mostly relies on identifying and labelling political entities who see their advertising rights constrained, trusted sources so TikTok can rely on trusted content, and content that is suspect but that the fact-checkers could not verify: *“we may add warning labels to content that our fact-checking partners cannot verify”*.

AIGC, for AI-Generated Content, is allowed to a certain point but under a label. TikTok relies on users labelling their own *“AI-generated or significantly edited content that shows realistic-looking scenes or people.”* Content that would fall into an already prohibited category is not allowed, like misleading about a natural disaster or conflict, harassing a public figure or *“Accounts focused on AI images of youth in clothing suited for adults, or sexualized poses or facial expressions”*.

In the *“Deceptive Behaviors and Fake Engagement”* chapter, it addresses anything that *“artificially boost engagement or trick the recommendation system.”* Under this metrics manipulation fall *“covert influence operations, impersonation, spam, fake reviews, and sharing hacked materials in harmful ways.”*

Those behaviours will result in an account ban and in banning *“additional or new accounts you create”*.

Covert Influence Operations are defined as follows: *“Coordinated, inauthentic behaviors where networks of accounts work together to mislead people or our systems and try to strategically influence public discussion. This may include attempting to undermine the results of an election, influencing parts of an armed conflict, or shaping public discussion of social issues.”*

Covert Influence Operations (CIO): Coordinated, inauthentic behaviors where networks of accounts work together to mislead people or our systems and try to strategically influence public discussion. This may include attempting to undermine the results of an election, influencing parts of an armed conflict, or shaping public discussion of social issues.

 NOT ALLOWED

- CIO activity, including:
 - Trying to deceptively influence elections, social issues, politics, or armed conflicts, including through hiding your account's real identity or intent
 - Accounts coordinating secretly to promote a candidate or political issue
 - Posting content on behalf of foreign entities (like a government or military) without saying so. We may make it harder to find content from accounts suspected of doing this while we review them
- Hacked materials distribution, when this poses a significant risk of harm. Note that we may allow limited discussion or distribution of hacked materials if there's a clear public interest and the content follows journalistic best practices
- Spam, such as:
 - Using automation to run many accounts or send repetitive content
 - Posting a large amount of irrelevant material
 - Buying or selling followers or engagement for financial gain
- Using bots or scripts to write fake reviews or comments, or to increase likes or shares
- Impersonation by pretending to be someone else without clearly stating that the account is a fan or parody account in the display name
- Pretending to be a fake person or organization with the goal of misleading people
- Circumvention of our policies, which includes:
 - Spreading violative content across multiple accounts
 - Using a different account to continue violating policies after being banned
 - Returning to TikTok after being permanently banned for **severe violations**
 - Using another account to avoid restrictions, such as comment blocks or FYF-ineligible content restrictions
- Trading, marketing, or providing access to services that artificially increase engagement, such as:
 - Followers or likes
 - Fake reviews
 - Using AI or bot accounts to drive traffic
- Sharing how-to guides or tips for boosting engagement in fake or deceptive ways

 FYF INELIGIBLE

- Tricking others into increasing engagement, such as:
 - "Like-for-like" promises
 - False incentives for gifting or following
 - Misleading claims meant to boost views or popularity

Figure 8. TikTok Community Guidelines, "Integrity and Authenticity", section "Deceptive Behaviors and Fake Engagement". [tiktok.com](https://www.tiktok.com) ↗

This indicates TikTok probably tracks known hacked material circulating on its platform (and then reviews if it circulates for legitimate or journalistic purposes), tracks any sign of automation (timestamps of account activity to detect coordination, copy-pasted or similar content), metrics manipulation and inauthentic behaviour in likes and reviews.

Multiple accounts are allowed: "You can have multiple accounts—for example, for fan content or creative expression", but all the above signs of coordinating them will result in a ban of the suspected account and linked accounts.

TikTok's page "Countering deceptive behavior" [tiktok.com](https://www.tiktok.com) ↗ also sheds some light on how they counter CIBs.

They prohibit (and check for):

“automation to register or operate accounts in bulk.”

“manipulating engagement signals to amplify the reach of certain content” (bots, scripts or other means to distribute content or interactions in bulk)

“presenting as a fake person or entity that does not exist (a fake persona) with a demonstrated intent to mislead others on the platform” (meaning they probably check the content / intent of any account suspected of “fake persona”)

“selling followers or likes, or providing instructions on how to artificially increase engagement”

Covert Influence Operations receive special treatment and a dedicated team: *“we have dedicated, full-time international trust and safety teams with specialized expertise across threat intelligence, security, law enforcement, and data science. These teams continuously pursue and analyze on-platform signals of deceptive behavior, as well as off-platform activity and leads from external sources.”*

Against Covert Influence Operations, TikTok works on behavioural signals, content and narrative.

“Accounts that engage in influence operations often avoid posting content that would be directly violative based solely on platform guidelines. That's why we focus on accounts' behavior and technical linkages when analyzing deceptive behaviors”

They check for signs of coordination: *“For example, they are operated by the same entity, share technical similarities like using the same devices, or are working together to spread the same narrative.”*

Signals that the actors try to conceal their location or use *“fake personas to pose as someone they're not.”*

Content that attempts to manipulate the public debate: *“Accounts are attempting to manipulate or corrupt public debate to impact the decision-making, beliefs, and opinions of a community. For example, they are attempting to shape discourse around an election or conflict.”*

They also look at *“off-platform activity [and] use open-source intelligence to identify any related deceptive behavior on TikTok”*.

Fake engagement and spam are separated from Influence Operations as being *“easier to spot”*, according to TikTok.

When the two mix, namely a classic case of a Covert Influence Operation using an established spam or scam network, then TikTok struggles with where to put it but assures its users that such cases will still be prohibited as ordinary scams: *“These financially motivated cases are not*

classified as covert influence operations unless they meet our influence operations criteria, since they don't share the same strategic goals, technical signals, or deceptive tactics."

DISCORD

The particularity of Discord is that it allows bots but not spam.

"Our automated system will flag bots it suspects are being used for spam or any other suspicious activity. The bot, as well as the bot owner's account, may be disabled as a result of our investigation. If your bot's code is publicly available, please remove your bot's token from the text to prevent it from being compromised." discord.com ↗

Discord community guidelines are divided into two sections: *"respect each other"* concerns mainly content and *"respect Discord"* is about spam and inauthentic behaviour. This second section is reinforced by a *"Platform Manipulation Explainer"* discord.com ↗

Spam is detailed under its three different forms: robot acting as a robot, human acting as a robot, robot acting as a human:

"Sending unsolicited bulk messages or interactions (or spam) is one of the most common ways the user experience is disrupted. Spam can be sent by automated accounts designed for this purpose (spambots), normal user accounts that manually execute spammy actions, as well as by user accounts modified to perform automated actions (self-bots)."

discord.com ↗

They also check for signs of automation at account creation, and for discussions or servers dedicated to facilitating platform abuse: *"This includes but is not limited to spaces distributing: spambots, server "raid" tools, account-creation tools, token generators, CAPTCHA-solving services, and other spam tools."*

Inauthentic engagement is detailed in the same part. In Discord's own words, *"Inauthentic engagement is closely tied to, but distinct from the issue of spam"*.

It is defined as follows: *"We will consider engagement to be inauthentic if it is purchased, induced (like promises of financial reward), fraudulent (for example, it involves fake accounts), or falsified (for example, simulating completion of Discord features to obtain rewards)"*.

It is explicitly forbidden to use *"join for join"*, *"invite reward"* as well as purchasing fake members or selling artificial engagement. Discord also fights against a particular trade (and treats it under its trade policies), the selling of *"custom server invite links (also called a "vanity URL")"*.

"Fake accounts" are explicitly forbidden and, as the examples tell us, widely used to boost metrics or member counts. Discord has to be subtle, as it allows pseudonymous accounts but forbids *"fake profile"*: *"We define a fake profile as an online identity for an individual or entity that*

does not exist. Some fake user profiles can be identified by their use of stolen or generated images as avatars. Impersonation means pretending to be a real individual, group, or organization.” discord.com ↗

In the “how to avoid violating these policies” section, Discord tells us what it is checking:

Messaging multiple unknown users

Modifying the Discord client (and other signs of automation)

Metrics inauthenticity, growth spikes

Signs or reports indicating the selling of a Discord account, username, server, server permissions or “vanity URL”

The policies against fraudulent coordination are in the Deceptive Practice explainer discord.com ↗

This policy details the scams, the malicious conduct (hacking, phishing, malware distribution, DoS attacks) and the “Fraud Services”. “We define fraud services as coordinated activities designed to generate fraudulent profits at others’ expense. Profit may include monetary gains or property obtained from third parties.”

Coordination generally suffers from poor coverage in the policies apart from this: “Do not promote, coordinate, or engage in harassment. We do not allow any type of harassing behavior, including sexual harassment, ban or block evasion, or coordinating server joins for the purpose of harassing server members, also referred to as “server raiding.”” discord.com ↗

Another big line of defence is a spam filter. It will check message similarity, frequency of posting, server joining rate or direct message activity:

“Joining a lot of servers simultaneously or sending a large number of friend requests might be considered spam. In order to shut down spambots, we take action against accounts that join servers too frequently or send out too many friend requests at one time. (...) if, for example, you send a friend request in just a few minutes to everyone you see in a thousand-person server, we may take action on your account.”

discord.com ↗

Finally, Discord has a “misinformation policy” dedicated to content analysis and covering “Health Misinformation” likely to result in harm: “anti-vaccination claims, dangerous and unsupported treatments for health issues, distortion of disease information”. It also covers “Civic Disruption”: i.e. “the posting, promotion, or organization of communities that share false or misleading claims that could disrupt or undermine the civic processes. (...) the promotion or organization of attempts to intimidate voters or activities that suppress civic participation.”

A civic process is defined as “any procedure run by governments or international political institutions that relies on public participation to make a decision or reach outcomes that impact

our society. Examples of civic processes include elections, referendums, and censuses.”

[discord.com](#) ↗

AMAZON

Although it is not the heart of its business model, Amazon still relies on authentic reviews and comments on the products. Therefore, it has even more incentive to counter CIB (and a whole dedicated policy against *“fake review brokers”*):

[trustworthys shopping.aboutamazon.com](#) ↗

Amazon relies on three main pillars to fight CIBs (in Amazon’s words: to provide a *“Trustworthy shopping experience”*) [trustworthys shopping.aboutamazon.com](#) ↗

Gate control: via seller identity verification and purchase history for reviewers

Human experts coupled with automated systems to check behavioural and content signals: *“combine advanced technology with human expertise to stay ahead of evolving threats. Multimodal systems analyze billions of signals simultaneously—from visual elements and textual content to seller behavior and supply chain patterns—revealing connections that isolated analysis would miss.”*

[trustworthys shopping.aboutamazon.com](#) ↗

Imposing costs on the attacker by working with law enforcement. Their teams *“work alongside law enforcement, brands, and industry leaders to combat organized retail crime, counterfeit operations, refund fraud, and scam networks.”*

Amazon is imposing a purchase history for platform activity. This implies each account is tied to a credit or debit card and creates a cost-of-entry rule.

“You must have spent \$50 on Amazon.com, using a credit or debit card, in the past 12 months, to: Create reviews (including star ratings); Answer customer questions ; Submit helpful votes ; Create idea lists ; Follow other contributors; Create posts on Amazon Inspire; Comment on Shop by Interest posts and reply to others' comments”

[amazon.com](#) ↗

Seller identity verification is even more severe: *“When sellers apply to sell in Amazon’s store, they must provide government-issued IDs, business credentials, bank statements, and proof of address”* (2025 Trustworthy Shopping Experience Report) and these documents are checked against Amazon’s document forgery detection systems and video verification.

Amazon is using an *“Account Health Rating”* to have a score and attribute a reputation to sellers' accounts, according to their compliance with Amazon Policies.

Amazon admits developing technology to assist human teams in the detection of behavioural signals and content analysis: *“we built systems that analyze details from across the selling and shopping experience—visual elements, textual content, seller identity, behavioral patterns, supply chain data, and network connections.”* These signals include customer behaviour on reviews, refund requests and customer service contacts: *“Customers share what they're experiencing with us in many ways—through reviews, refund requests, and customer service contacts.”* They are also developing AI in behaviour detection (in their own words *“AI-based systems that reveal patterns that would otherwise not be discernible”*) (2025 Trustworthy Shopping Experience Report).

Another defence is to use language borders. They only allow comments and reviews in a certain language if it matches the country: *“We only allow content to be written in the supported languages of the Amazon site where it will appear. For example, we don't allow reviews written in French on Amazon.com. It only supports English and Spanish.”* Comments in a mix of languages are not allowed. amazon.com ↗

It is a rough filter: writing in the language of the targeted country is the first thing any CIB does, commercial or political (unless an operator forgets to delete the Urdu prompt), so the rule costs attackers little; but it shows that Amazon detects inauthentic language, including mixed languages.

External links are simply banned.

Posting any content that would include a phone number, email address, mailing address, licence plate or order number is forbidden. It is not difficult for automated systems to go fishing for these standardised formats.

On the content side, profanity and hate speech are forbidden. Discussing *“sex and sensuality products sold on Amazon”* is allowed, as is discussing products with sexual content like books or movies, but *“profanity or obscene language”* is not.

Under *“profanity, harassment”* there is a strict rule against coordination as a whole. *“Don't post from multiple accounts or coordinate with others”*. amazon.com ↗

Finally, spam filters will look for *“repetitive text, nonsense and gibberish, content that's just punctuation and symbols”* and ASCII art (probably because it is difficult to train anti-spam to see the difference between ASCII art and *“content that's just punctuation and symbols”*).

The hard work for Amazon is to sort, in the content, the authentic from the inauthentic reviews. It is forbidden to:

“Create, edit, or remove a review in exchange for compensation, including payments, refunds, discounts, products, gift cards, warranties, or services. [...] Post reviews on your own product (if you’re also a seller) or on a competitor’s product. Post reviews on products that you have financial interest in. Post reviews on products from sellers, authors, brands, or artists that you have a personal connection with. [...] Post a review for a product before it has been delivered to you.”

These can be easily checked by comparing the review timestamps with the delivery status, refund events followed by review edits or removal.

Review manipulation is carefully watched and a whole part of the policies is dedicated to preventing manipulation: “Anti-Manipulation policy for Customer Reviews”, and Amazon even threatens lawsuits:

“We pursue lawsuits for reviews manipulation against dishonest sellers and manufacturers who attempt to purchase fraudulent reviews and the parties who provide and post those reviews. These lawsuits have produced monetary judgments exceeding the annual revenue for such sellers and data allowing us to take additional enforcement actions against others.”

[amazon.com](#) ↗

Fake reviews and “review abuse” are tracked by both automated means and expert investigators: “Amazon aggressively fights fake reviews, using a combination of machine-learning models along with expert investigators to ensure that every review in our store is authentic”.

[trustworthyshopping.aboutamazon.com](#) ↗

AI is trained and used to detect authenticity: “before a single review appears in our store, artificial intelligence examines thousands of data points simultaneously—account relationships, sign-in patterns, review history, behavioral anomalies. Machine learning models work alongside large language models and deep graph neural networks to detect patterns humans cannot see.”

Amazon also relies on AI to detect false positives, which is more worrying:

“A product accumulating reviews quickly might simply be great. Poor grammar in a review doesn’t necessarily signal fraud. Telling the difference between authentic reviews and abusive or fake ones requires detecting patterns that only become visible through AI-powered analysis of billions of reviews built up over nearly three decades.”

[trustworthyshopping.aboutamazon.com](#) ↗

Amazon does not separate commercial CIBs from political CIBs. Any form of coordination is prohibited and falls under the same rule. Coordination is tracked by an AI system named “Knowledge Graph Technology”: “Traditional fraud detection examines signals one at a time: Is this account suspicious? Does this review look fake? But bad actors have learned to game these systems. A review might look authentic on its own, but when you map its relationship to other

accounts and behavioral patterns, a network of coordinated abuse emerges.” Amazon uses Knowledge Graph Technology to detect suspicious coordinated behaviour. “Deep graph neural networks uncover signals through behavioral patterns, detecting coordinated abuse that would be difficult to detect if each signal were evaluated on its own—whether that’s fake review rings, counterfeit networks, refund fraud schemes, or scam operations.”

This technology is greatly facilitated by the fact that Amazon simply does not have to sort between authentic and inauthentic coordination: any form of coordination on Amazon is suspect.

CIBS AND AI

Automating a CIB network before AI was a complex procedure. It required investing in technical capabilities and meant a higher cost for the attacker when the network was identified and disabled. It was at the time much more efficient and cheaper to invest in a human click farm. This cost has considerably decreased, rendering CIB automation within reach of considerably more bad actors. The AI agents are used to hide behavioural signatures and LLMs to generate original content that is more difficult to detect with copy-paste or similar-content detection.

Even if the old techniques remain highly efficient with a small update, in-depth analysis of the content and the narrative is now an obligation.

AI CAPABILITIES OFFERED TO ADVERSARIES

Even Meta, the champion of the content / behaviour separation, reluctantly admits that CIBs boosted with AI require some adaptation:

“our core assessment from 2025 holds: behavioral detection approaches remain effective against AI-enabled influence operations. The networks disrupted during this period were identified through behavioral signals, technical indicators, and network analysis—signals that are difficult to mask entirely, even with AI.

However, the velocity of adoption and integration we observed warrants continued investment in AI-enabled defensive capabilities. The industrialization of AI workflows, in particular, suggests that the next phase of this competition may be defined less by content quality than by operational scale and adaptation speed.” (Meta, [Adversarial Threat Report, H2 2026](#))

In concrete cases, like the fight against romance scams augmented with AI, Meta openly admits to doing content analysis: *“We used specialized techniques to detect and analyze the content and behavior of these scams, which enabled us to find links across the network. We have used these signals to enhance our detection capabilities associated with AI-assisted scam operations.”* (Meta, [Adversarial Threat Report, First Half 2026](#))

The shift in methodology is huge. Initially, Meta’s policy toward CIB was explained by Nathaniel Gleicher using the *“haystack / needle”* metaphor: to use automated systems working on behavioural signals to *“shrink the haystack”* and human overview to *“find the needle”*.

"We combat coordinated inauthentic behavior in two ways. First, our expert investigators use skills brought from the worlds of cybersecurity research, law enforcement, and investigative reporting to find and take down the most sophisticated networks. To do so, they collaborate closely with our data science team, which uses machine learning and other advanced technologies to identify patterns of malicious behavior. Second, we build technology to detect and remove automatically the most common threats. If expert investigations are looking for a needle in a haystack, our automated work is akin to shrinking that haystack. It reduces the noise in the search environment by removing unsophisticated threats."

(Nathaniel Gleicher, Head of Cybersecurity Policy, Facebook, written testimony before the US House Committee on Oversight and Reform, Subcommittee on National Security, 22 May 2019)

Meta is now doing quite the opposite: a fine analysis of content and behavioural signals to train automated systems. Humans find the needle by combining content analysis and behavioural signals; trained robots are then sent to explore the haystack.

AI CAPABILITIES OFFERED TO THE DEFENDERS

This shift in methodologies, employing a more hybrid approach, is made possible because new AI capabilities are also offered to the defenders. They can be used to detect more behavioural signatures but also to help content and narrative analysis at a scale previously unknown. These tools also benefit from an agnostic and neutral image in the eyes of the public and the engineers. Narrative analysis from a human is still suspected of bias, while narrative analysis from an AI will for a time be believed to be neutral. Anthropic's Clio is the model of this: a system that reads millions of Claude conversations, groups the similar ones and shows its analysts only anonymised topic clusters, so that content analysis can be presented as *"privacy-preserving"* because no human reads the content (anthropic.com ↗).

These two factors allow platforms to include AI in their content analysis, to boost their anti-spam with new content analysis capabilities, as Amazon's *"knowledge graph technology"* shows.

Now platforms can use AI to look at the content en masse. Meta now uses AI to analyse content in its fight against cartels and drugs. Even more telling: AI is doing content analysis to help human experts working on the behaviour: *"AI-Driven Concept Detection provides significant advantages in the discovery phase of Strategic Network Disruptions, particularly to expert teams that are already deeply familiar with cartel behaviors."* (Meta, Adversarial Threat Report, First Half 2026)

Clio is also how Anthropic finds coordination upstream:

"Clio has proven effective at identifying patterns of coordinated, sophisticated misuse that would otherwise be invisible when looking at individual conversations [...]. For example in late September, we identified a network of automated accounts using similar prompt structures to generate spam for search engine optimization. While no individual conversation violated our Usage Policy, the pattern of behavior across accounts revealed a form of coordinated platform abuse we explicitly prohibit in our policy and we removed the network of accounts."

(Anthropic, "Clio: A system for privacy-preserving insights into real-world AI use", 12 December 2024)

AI COMPANIES' THREAT REPORTS

AI companies themselves have a complete disregard for the frontier between content analysis and behavioural signals. Our previous version of the detection tree had included the *"intent detection"*, a technique that can only be done by looking at the content. Intent detection is for instance largely used by AI companies in their fight against CIBs, as they carry out large-scale prompt analysis. OpenAI's threat report clearly used a combination of the two to identify a cluster of Chinese accounts. AI companies cannot really work around analysing their users' prompts to identify the intent.

Actor

We banned a cluster of ChatGPT accounts that likely originated in China and used ChatGPT to generate social media content for a covert influence operation. They prompted ChatGPT in Simplified Chinese while repeatedly asking for English- and Chinese- language outputs, posing as Americans from a variety of backgrounds, that were posted across multiple social media platforms. As we do not allow access to our models from China, they used VPNs to access our platform.

The operators of the accounts were likely part of a social media operations team at a private Chinese technology company conducting work for Chinese provincial-level government clients. This activity appears consistent with a commercial ecosystem that supports Party-state priorities in public opinion guidance. A separate report they uploaded to ChatGPT described their objectives and strategies for influencing public opinion and establishing social media accounts designed to evade platform detection systems.

Figure 9. OpenAI, June 2026 Threat Report, "PRC-linked influence operations are targeting AI debates in the US", section "Actor". cdn.openai.com

The prompt and content analysis is likely to become a standard in CIB detection. AI models in CIB are primarily used to generate fake identities and generic content that will be posted on social media. Therefore, AI companies have to work on cross-platform detection and will more and more team up with the targeted platforms.

OpenAI, in a chapter called *“platform operation”*, explains:

“The accounts asked ChatGPT to generate, polish and edit work reports that revealed the operational security considerations of their activities on social media and their understanding of platform detection systems. They described their objectives to include establishing persistent and credible accounts, producing visually appealing content to expand audience reach in target regions, and maintaining long-term account viability by anticipating platform enforcement”

This does not mean that behavioural indicators are irrelevant. They are still invaluable and heavily used by AI companies to identify threat actors, especially, but not only, when the operation goes live on social networks.

In its October 2025 report, in a cyber case attributed to North Korean IT workers rather than an influence operation, OpenAI tracked *“accounts engaged with our models primarily in the Korean language, showing structured workflows with many accounts active in narrow time windows.”* This means OpenAI looked at the language behaviour combined with narrow time frames. But it has to be combined with prompt analysis: *“Each of these accounts appears to have focused on a specific use case, for example converting Chrome extensions to Safari for Apple App Store publication, configuring Windows Server VPNs, or developing macOS Finder extensions - rather than each account spanning multiple technical areas”* cdn.openai.com ↗

In another instance, OpenAI clearly identified threat actors based on behavioural signals:

“This network consisted of ChatGPT accounts that operated in a time pattern consistent with mainland Chinese business hours, prompted our models in Chinese, and used our tools with a volume and variety consistent with manual prompting, rather than automation. In one instance, we believe the same account may have been used by multiple operators.”

cdn.openai.com ↗

But the behavioural indicators are now used in nearly total fusion with content analysis.

Anthropic analysed an operation using a perfect combination of prompt analysis, content analysis and behavioural signals:

"We discovered that the operation repeatedly relied on three manipulation tactics: rewriting the same source story in opposite ideological directions for different audiences, adding political angles to stories that originally had none, and laundering stories across borders into unrelated regions, stripped of their original context."

(...) "The network paired each fake outlet with an X (formerly Twitter) account. These sites were then amplified by a layer of commenting accounts created during the exact same timeframe as the websites, with many using AI-generated profile photos. Most of these fake accounts were created in June and July 2025."

AI is used to generate content but also to operate the bots, automate coordination, hide inauthentic signals: *"Users are starting to use frontier models to semi-autonomously orchestrate complex abuse systems that involve many social media bots."* [anthropic.com](https://www.anthropic.com) ↗

AI labs are tracking coordination and behavioural signals accumulation to make sure they do not catch too many false positives: *"To detect and disrupt threats effectively without disrupting the work of everyday users, we employ a nuanced and informed approach that focuses on patterns of threat actor behavior rather than isolated model interactions"* cdn.openai.com ↗

These AI companies are working together with other platforms to counter CIBs, further increasing the harmonisation of the narrative / behaviour hybrid methodologies. OpenAI's usage policies carry this changelog entry: *"2021-02-26: Clarified the impermissibility of Tweet and Instagram generators."* openai.com ↗

CONCLUSION

This chapter is the first of a 5-chapter series on CIB detection.

Here we tried to read in the platform policies how they define and detect Coordinated Inauthentic Behaviours. Although the techniques themselves can't be fully reproduced by researchers, three findings stand out.

First, the barrier between behaviour and content has fallen. The separation that Facebook wrote into its doctrine from 2017 (*"based on their behavior, not the content they post"*) and that was pushed on the analysts by some organisations as a promise of neutrality no longer holds, by the platforms' own admission: the same platforms that restate *"not the content posted"* concede that they combine *"behavioral and content signals"*, write content-quality rules into their spam policies, and let automated systems read narratives. Bluesky, LinkedIn, YouTube and X have even merged their *"authenticity"* rules and their content rules under one heading.

Second, AI changes the economics of CIB, not its techniques. Original text defeats copy-paste detection and agents can hide a posting cadence, but the networks are still caught by creation dates, shared infrastructure, recycled personas, prompts pasted into posts and watermarks left on images. The AI companies, now the upstream tier of detection, detect on behaviour and prove with content: the prompt is the intent.

Third, platforms did not cross this line on their own. The 2024 election year, the Code of Conduct on Disinformation, the Commission's guidelines asking for one team that spans content moderation, fact-checking and threat disruption, and the Digital Services Act pushed them to write the fight against influence operations into their terms, and to politicise rules that were designed to be content-agnostic.

For the researcher the consequence is practical. The detection tree can keep its 2021 organisational branches, even if every branch needs to be reviewed and updated, with a refined list of what each detection method can catch. These chapters will be published in the weeks to come.

Antoine de Gunzbourg for MELETA.