Guarding against Algorithmic Amplification of Synthetic Information

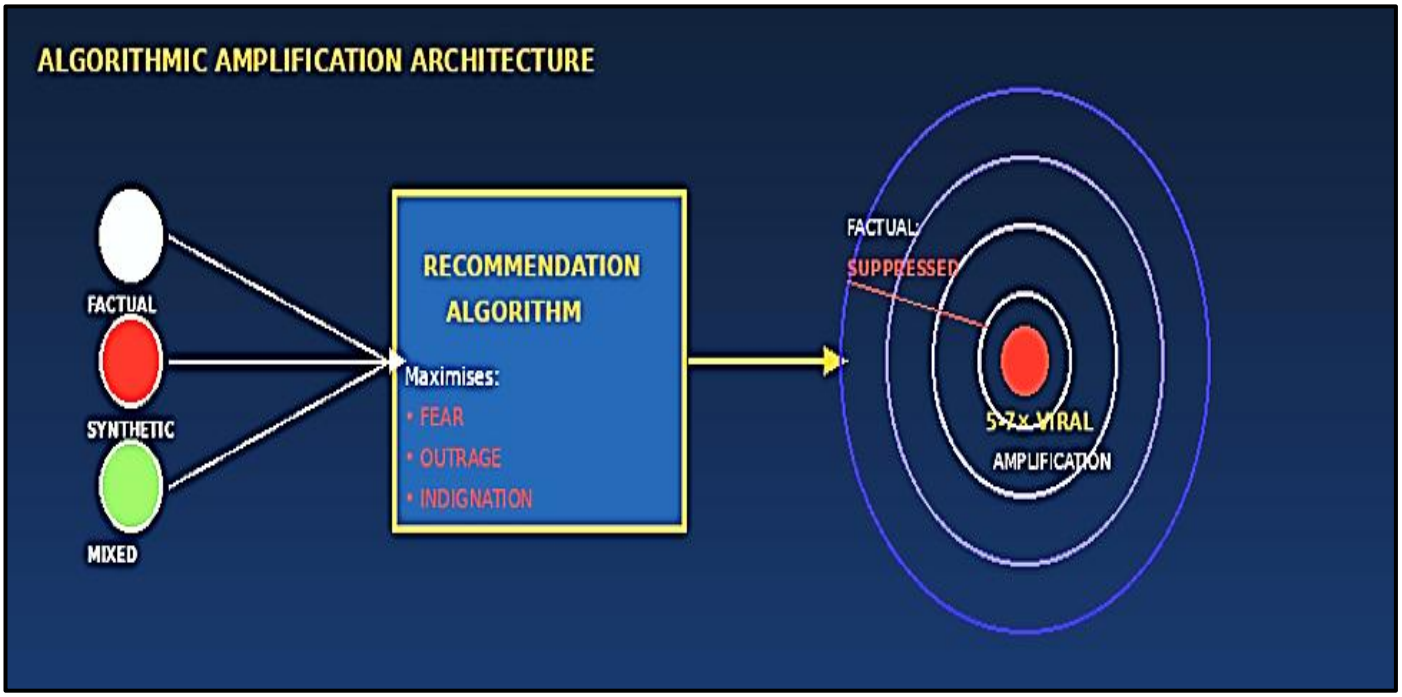
Preamble

It is not just the creation of the synthetic information but rather the amplification of such information that presents the strategic challenge. Even when a sophisticated deepfake or synthetic information reaches just a few hundred individuals, there is very little strategic value in that piece of information. However, the same piece of information, if used for malicious purposes via recommendation algorithms, will reach tens of millions within a few hours and change perceptions regarding military activities and politics as well as national security scenarios.

Ten-Point Advisory

1. **Understand the Architecture of Algorithmic Amplification**.

Algorithms used by social media companies such as YouTube, Facebook, X (previously Twitter), Instagram and Telegram are designed for maximising engagement irrespective of what is true. Content that incites fear, outrage and moral indignation yields higher engagement rates than facts, leading to additional amplification by algorithms. According to studies, false or sensational information gets up to five to seven times more algorithmic boost compared to news stories.

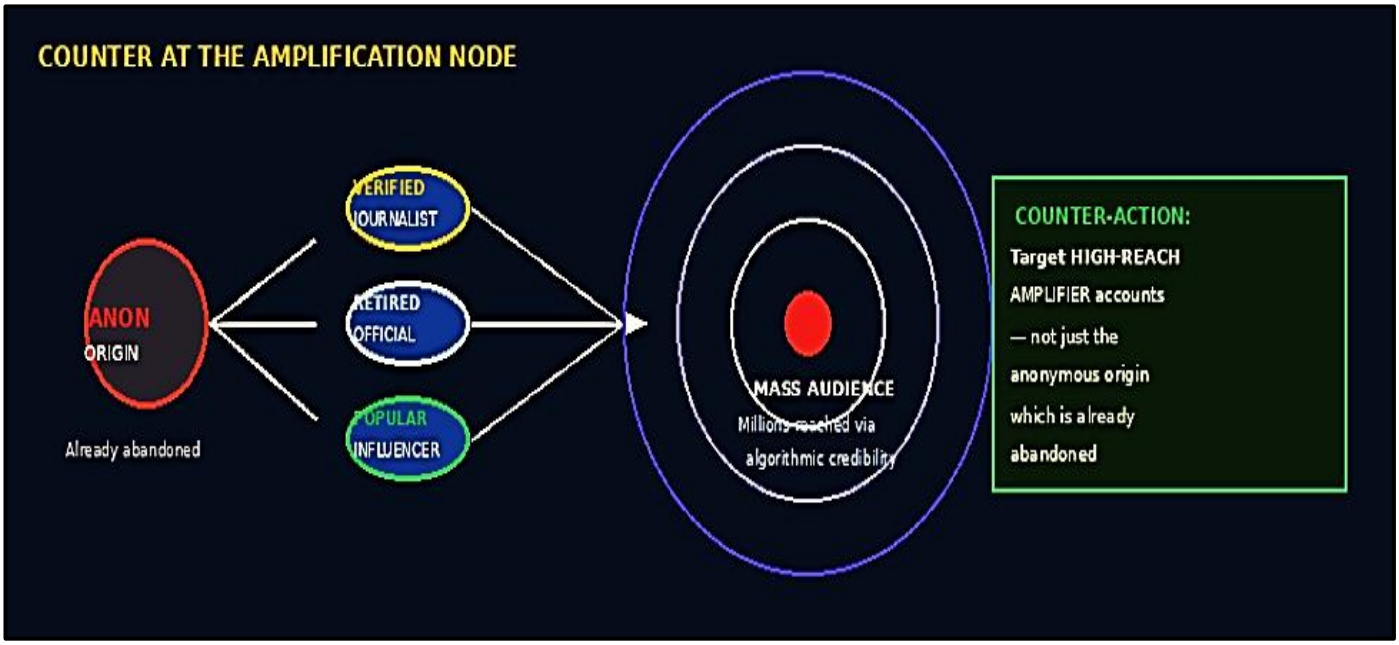


2. **Interrupt the First 30 Minutes of Virality**.

Disinformation attains maximum traction within the first 30 minutes of being posted. A synthetic video which is able to acquire critical momentum of re-shares within this time frame obtains the social validation that is hard to undo. The platform's anti-disinformation mechanisms like friction labels, share-throttle measures, algorithmic downgrading are only successful if they are applied within this crucial period. Surveillance systems should be able to detect emerging disinformation videos within this 30-minute period.

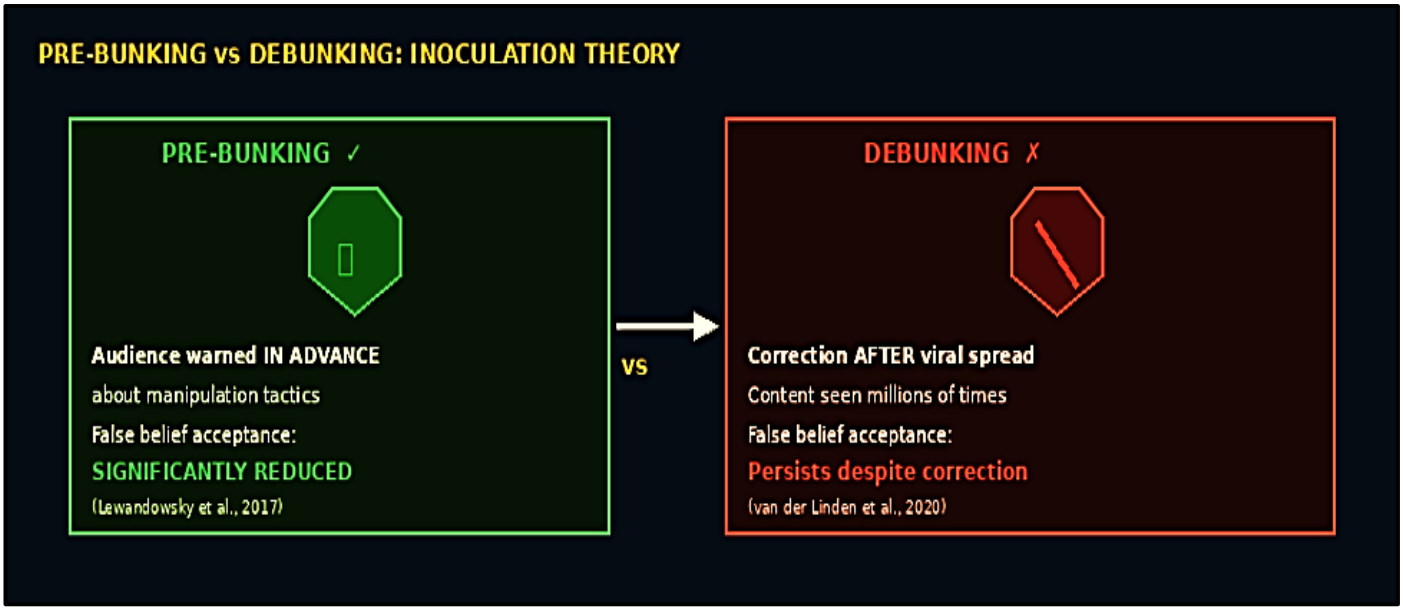
3. **Counter at the Amplification Node, Not Just the Origin**.

Adversaries plant fabricated content via reputable intermediary accounts i.e. reporters, former senior government officials or popular social media users; to take advantage of their algorithmic trust. In response, it is imperative to target the amplification nodes that help spread fabricated content rather than just targeting the anonymous account from which the content originated.



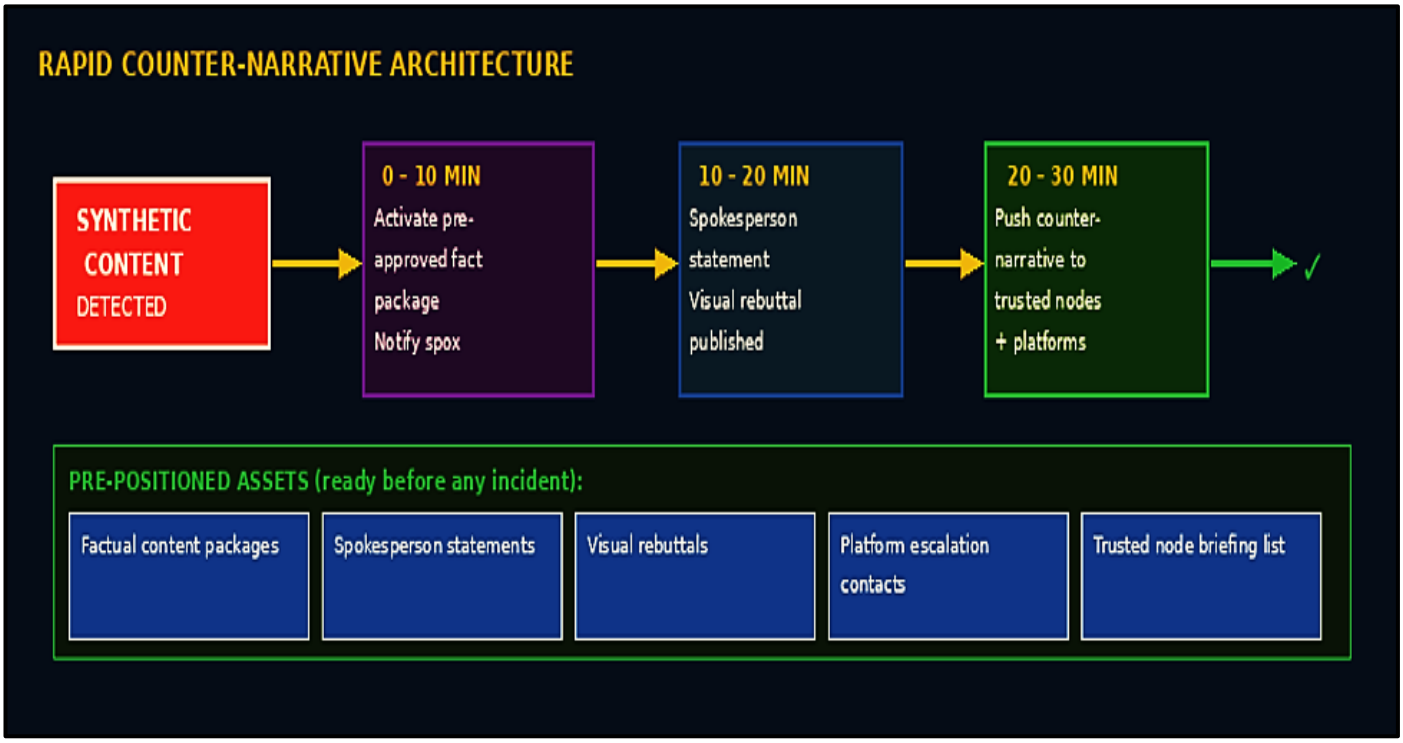
4. **Deploy Pre-Bunking Before Debunking**.

A study conducted by Lewandowsky, Ecker and Cook (2017), along with another one carried out by van der Linden et al. (2020), proves that pre-emptive inoculation, which entails warning an audience of the possible manipulative tactics employed against them before their deployment, is far more efficacious than correction after a manipulation technique goes viral. Pre-bunking is shown to be more effective than corrective communication in reducing the uptake of false beliefs by a large margin.



5. **Establish a Rapid Counter-Narrative Architecture**.

Any synthetic message which reaches viral velocity needs to be met with a counter-narrative just as fast and just as credible. Always have ready pre-vetted packages of facts, statements from the spokespersons and visual arguments, that can be released 30 minutes after the synthetic message becomes viral. It is speed and credibility which determine the success of the counter-narrative.

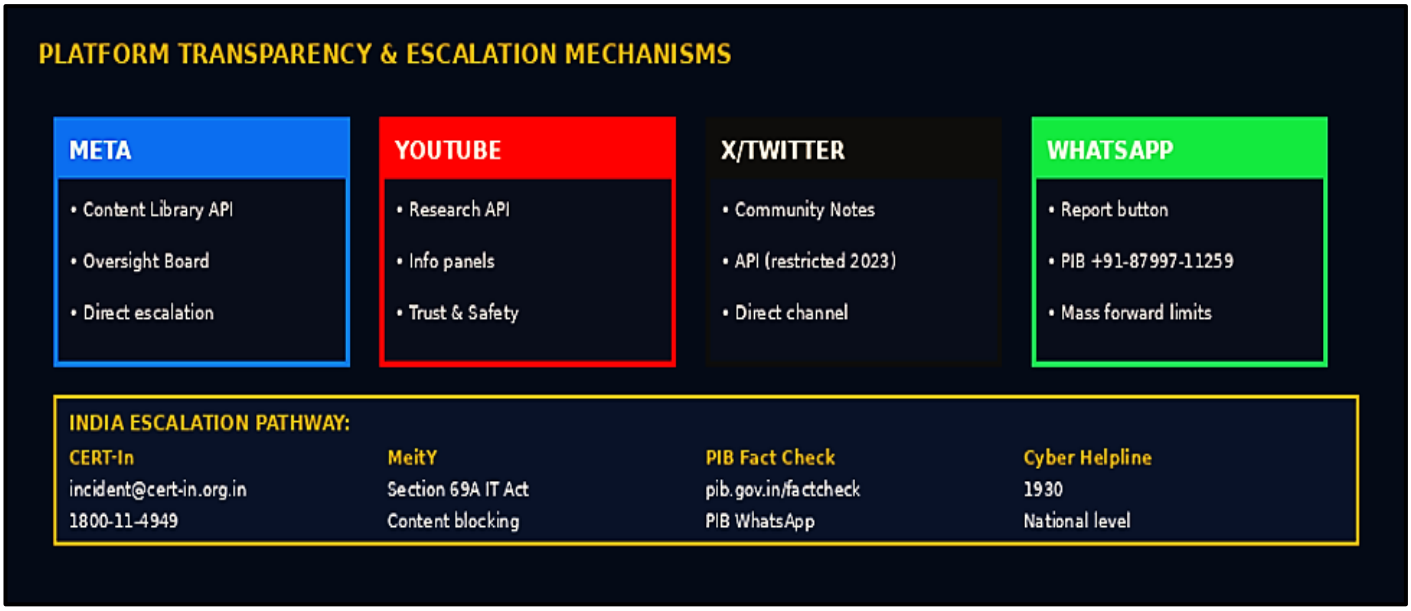


6. **Engage Platform Transparency and Escalation Mechanisms**.

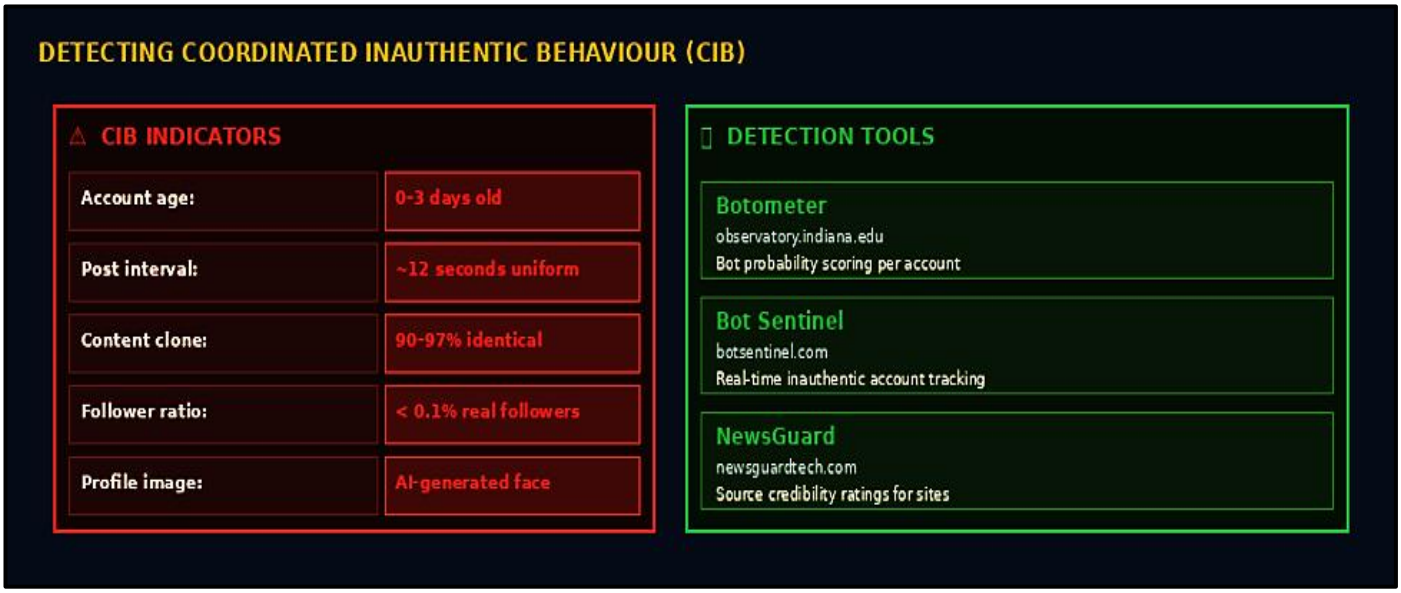
Utilise platform APIs and escalation systems: Content Library API from Meta, YouTube Research API and direct trust and safety escalation system. Note: X (formerly known as Twitter) has drastically reduced API accessibility to researchers in 2023; hence, the Academic Research API is very rarely accessible. Collaborate with CERT-In (incident@cert-in.org.in / 1800-11-4949) and MeitY in case Section 69A of IT Act needs to be used for synthetic content that has national security implications.

7. **Recognise and Counter Coordinated Inauthentic Behaviour (CIB)**.

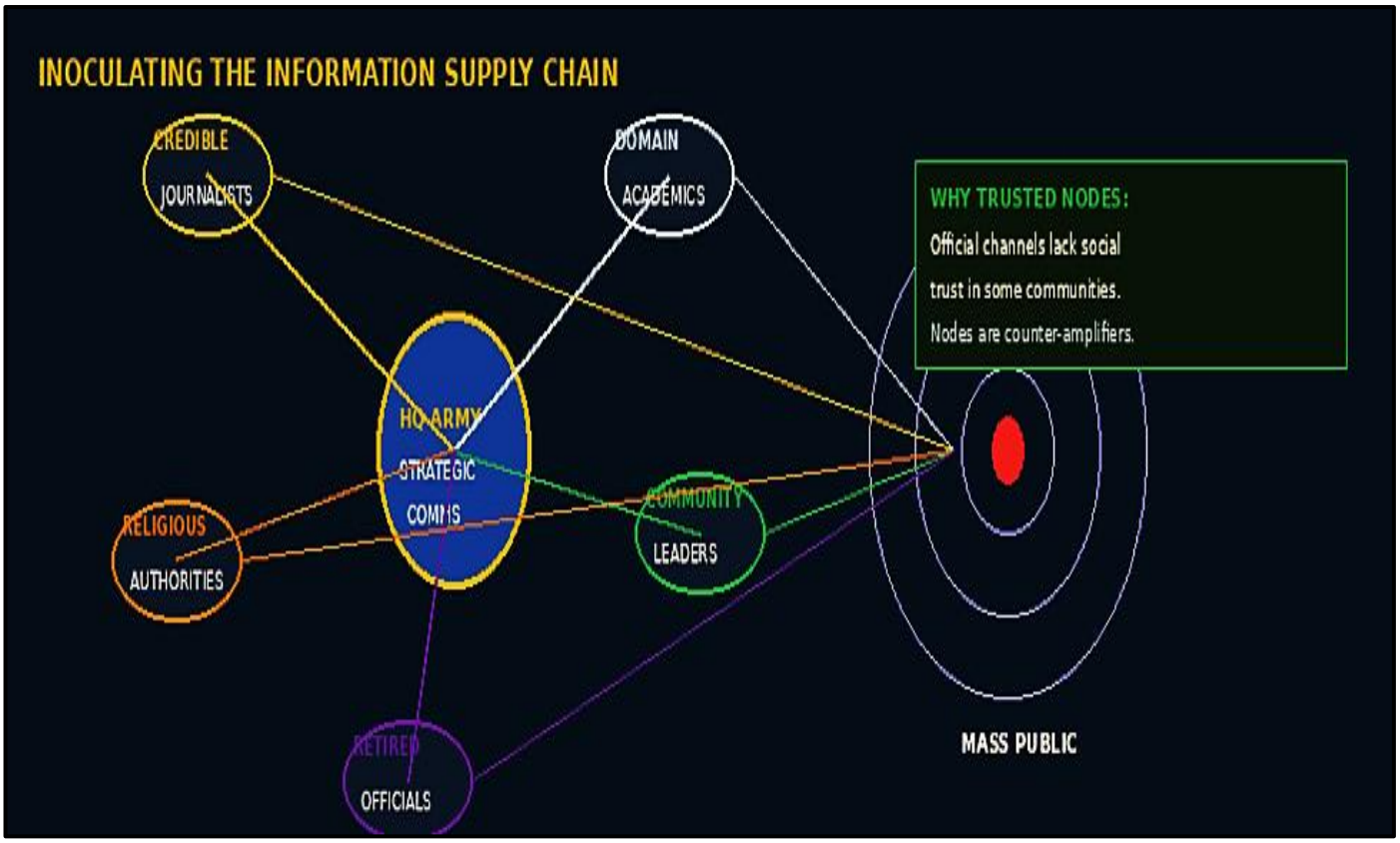
The amplification of synthetic material often relies on coordinated networks of inauthentic accounts like bot farms, troll armies or sock puppets which generate false signals. Features that point to CIBs: age of the account being between 0 and 3 days; equal posting timespans of 12 seconds; similarities in the material posted ranging from 90 to 97%; followers per account under 0.1%; AI-created avatar images. Detection tools are available at Botometer (observatory.indiana.edu) and Bot Sentinel (botsentinel.com). NewsGuard (newsguardtech.com) now provides rating service instead of others.



8. **Inoculate the Information Supply Chain Through Trusted Nodes**.

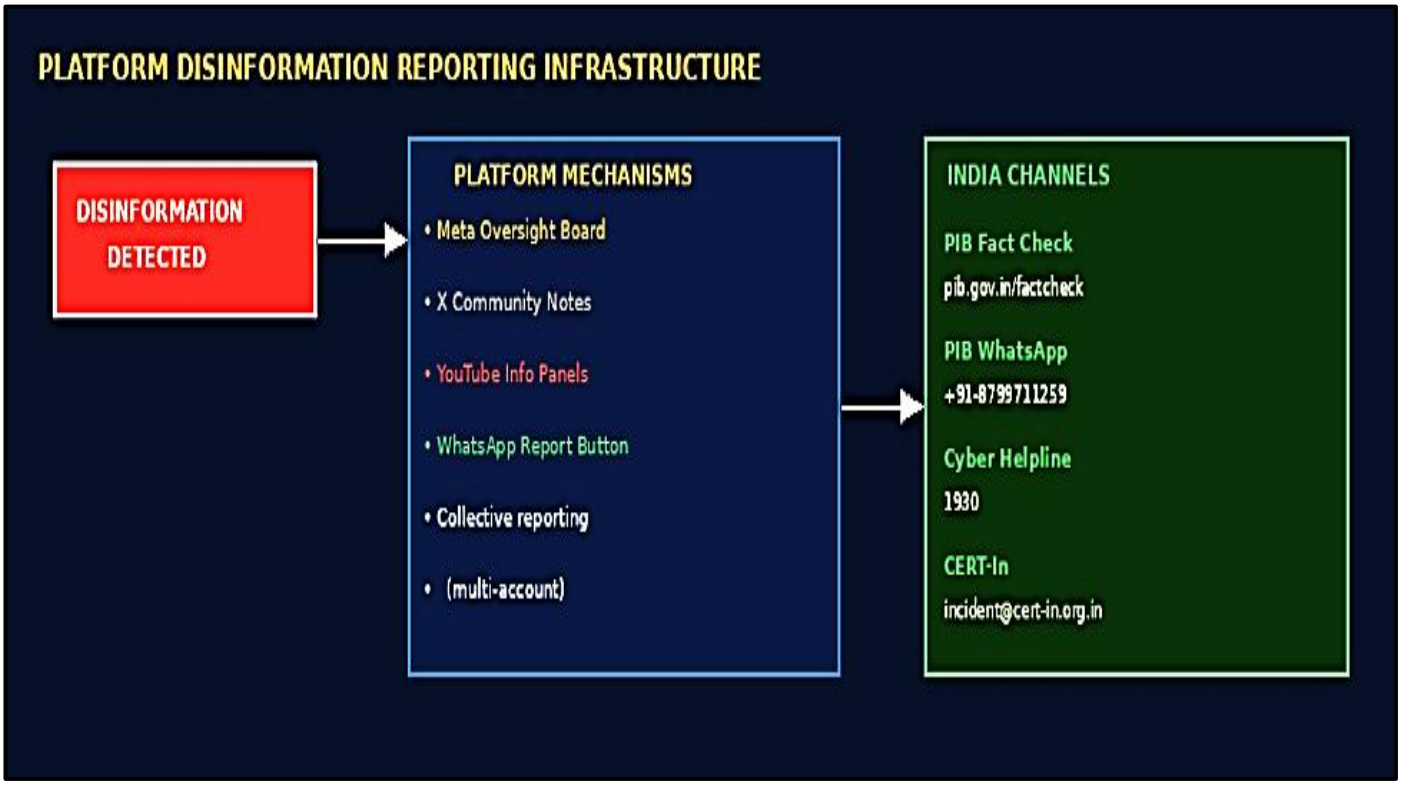


Identify and debrief credible messengers for information dissemination – journalists, academics from the relevant field, community members, clerics and former government officials who have enough clout in society – with factual information before the synthetic story gets disseminated completely. These credible messengers are used as counter-amplifiers within societies where the government is not sufficiently trusted to deliver information.



9. **Leverage Platform Disinformation Reporting Infrastructure**.

Use the platform-based reporting process: Meta's Oversight Board, Community Notes by X, information panel by YouTube and the report option within WhatsApp. Engage in group reporting from various reliable handles to escalate issues up to the platform level for intervention. Official channels for India include PIB Fact Check website (pib.gov.in/factcheck), PIB WhatsApp (+918799711259), Cyber Helpline (1930) and CERT-In. Do note that the MyGov number 7290006345 pertains specifically to the COVID situation and not disinformation.



10. **Build Institutional Digital Resilience Through Media Literacy**.

A lasting way to prevent the use of algorithms to amplify the reach of synthetic content is by having a literate populace that is able to critically evaluate information. There should be investments made in media literacy programs that focus on demographics that are especially vulnerable to such algorithm manipulation, including older individuals, new users in rural areas and politically divided communities.