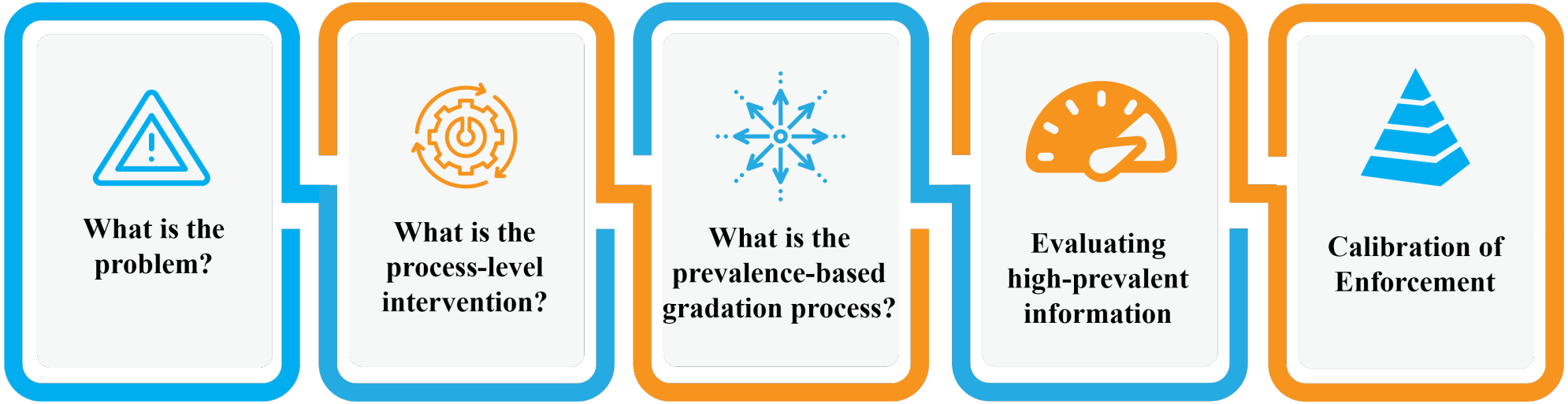
22 January 2023

A New Process To Tackle Misinformation On Social Media: Prevalence-Based Gradation

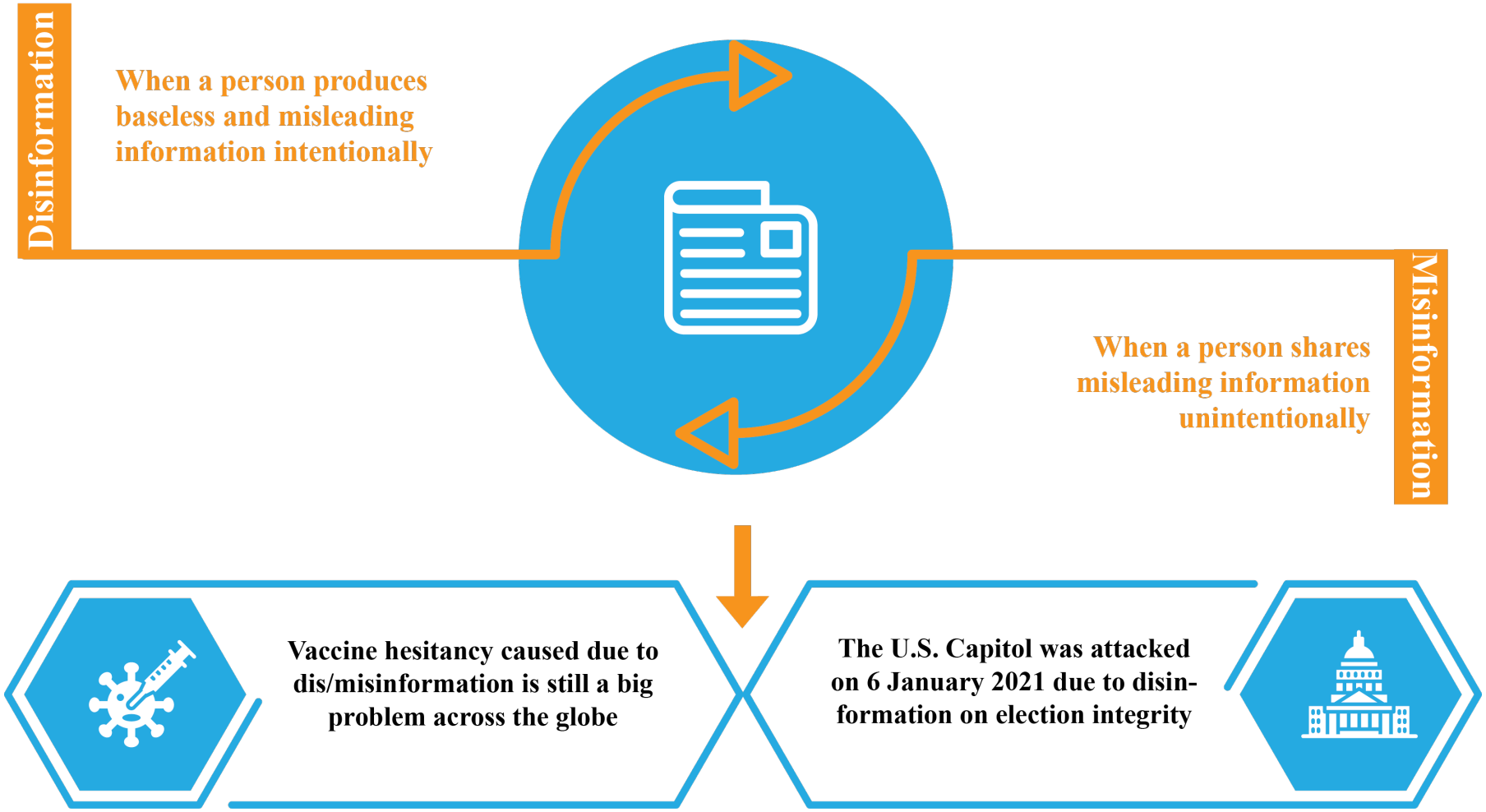
Enigma 2023

Kamesh Shekar
Programme Manager – Privacy & Data Governance Vertical, The Dialogue
kamesh@thedialogue.co

Outline


What is the problem?
What is the process-level intervention?
What is the prevalence-based gradation process?
Evaluating high-prevalent information
Calibration of Enforcement

What is the Problem? (1/2)



What is the Problem? (2/2)

- Every stakeholder starting from the government, social media platforms, and consumers, has a role in controlling the spread of misinformation and disinformation.
- But the role of social media platforms in tackling information disorder needs greater attention as they pose both competencies as well as the potential for causing ill effects.
- The platforms use various forms of technological measures like word filters, automated hash-matching and other predictive machine learning tools for detecting and tackling unlawful content such dis/misinformation related to vaccines, elections etc.

> **This illuminates that though social media platforms institute ex-post technological measures, we increasingly see content falling through the crack due to delayed responses, lack of efficient use of resources and false negatives. One of the critical reasons for falling through the crack is confinement to content-level intervention in the absence of process-level intervention. Therefore, to be proactive and not just reactive, we need robust ex-ante measures and means to implement the same efficiently.**

What is the Process-level Intervention?

Prevalence-Based
Gradation Process

High prevalent information

Corrective action

Bucket 1

Bucket 2

Bucket 3

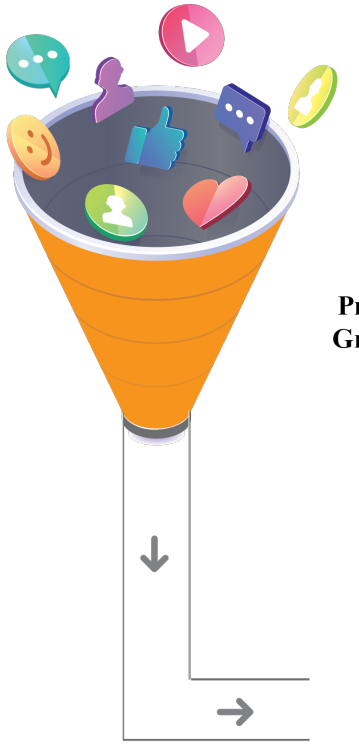
Bucket 4

Evaluating the prevalent information: Ex-ante scrutiny

Enforcing responsive moderation: Calibration of enforcement

Decision Communication

What is the Prevalence-based Gradation Process?

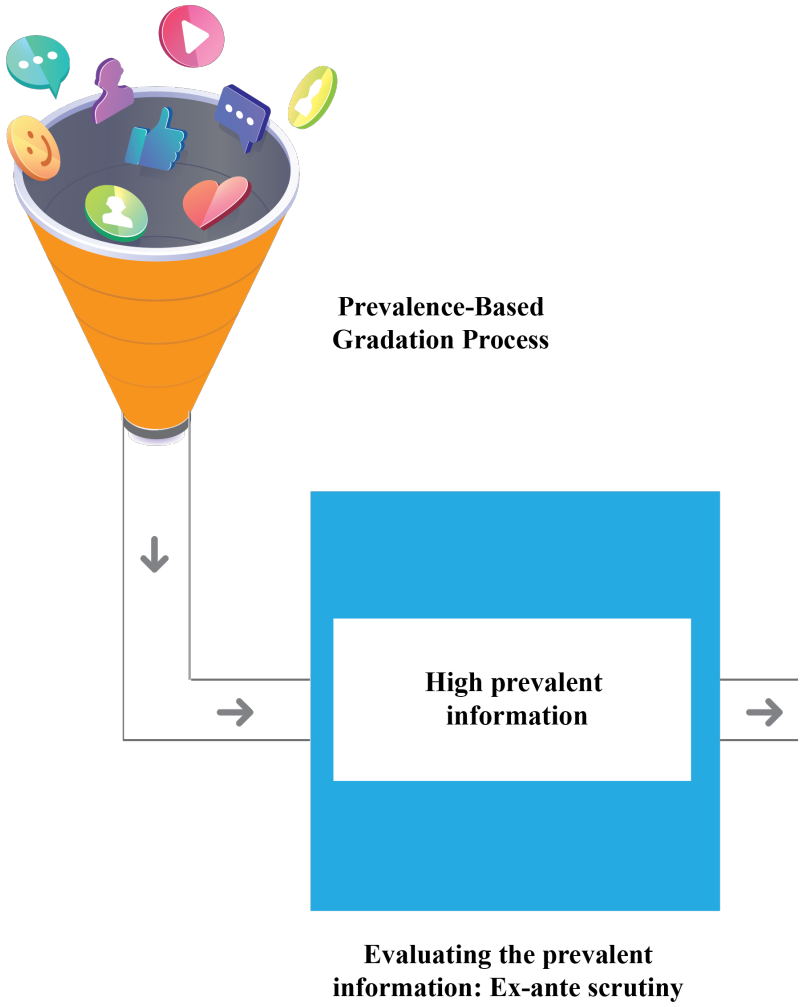


Prevalence-Based Gradation Process

User profile / High-stake information		No. of likes, shares and comments		
		High	Medium	Low
No. of followers	High	Bucket 1 – Severely high prevalent	Bucket 4: Chances for severely high prevalence	Bucket 7: Fewer chances for severely high prevalence
	Medium	Bucket 2 – Very high prevalent	Bucket 5: Chances for very high prevalence	Bucket 8: fewer chances for high prevalence
	Low	Bucket 3 – highly prevalent	Bucket 6: Chances of high prevalence	Bucket 9: no chance for prevalence

- The prevalence-based gradation process is a means through which social media platforms can evaluate content using ex-ante measures and exercise optimal corrective action in calibrated format adjusted according to the exposure level of the information.
- This process is performed by utilising the data collected from the users to bucket the information within the gradation matrix.
- This process should not be costing platforms much as they already capture data related to prevalence, reach etc.

Evaluating High-prevalent Information (1/2)

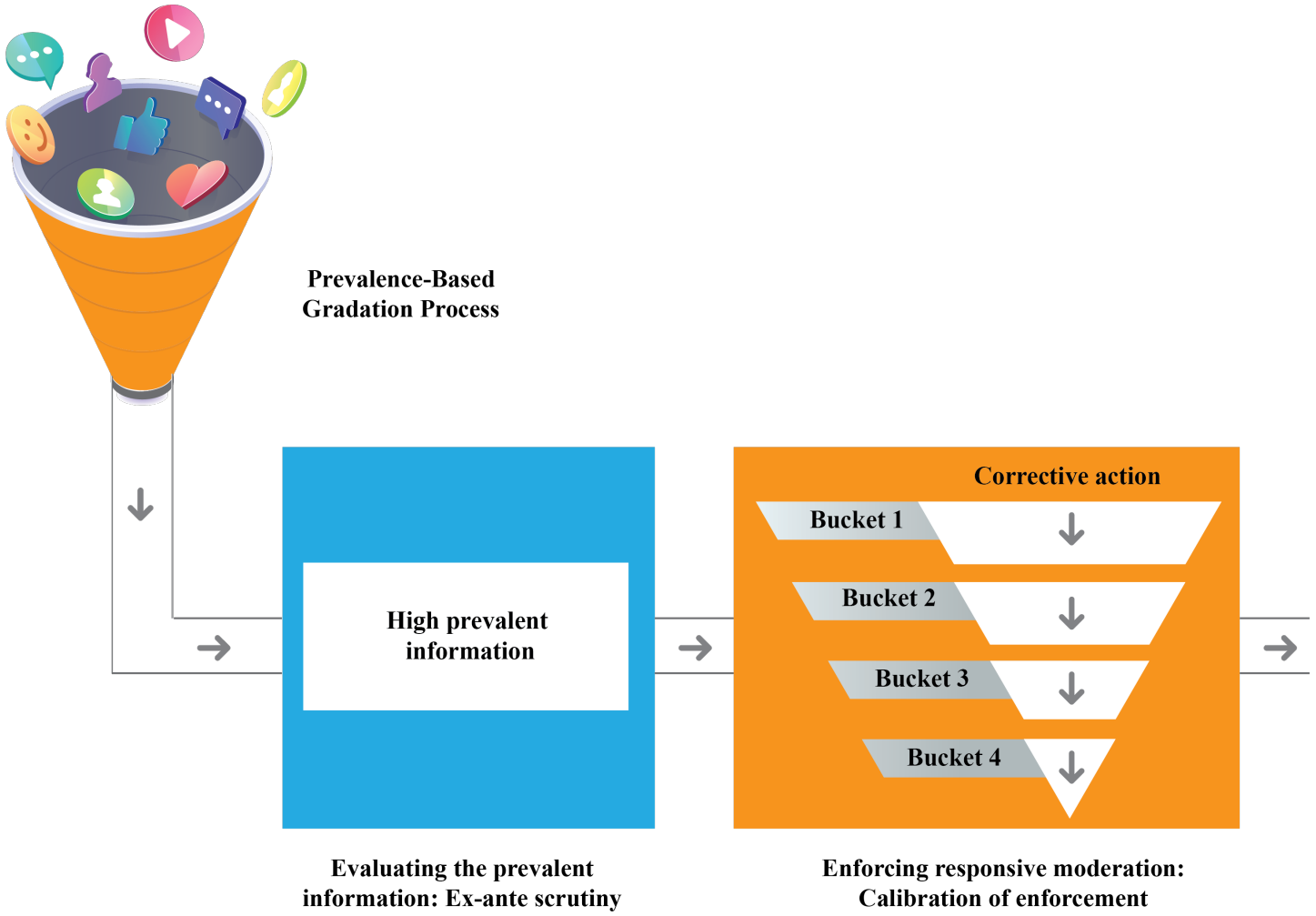


Evaluating High-prevalent Information (2/2)

User profile		High-stake information: High	High-stake information: Medium	High-stake information: Low
		No. of likes, shares and comments		
No. of followers	High	Bucket 1 – Severely high prevalent	Bucket 4: Chances for severely high prevalence	Bucket 7: Fewer chances for severely high prevalence
	Medium	Bucket 2 – Very high prevalent	Bucket 5: Chances for very high prevalence	Bucket 8: fewer chances for high prevalence
	Low	Bucket 3 – highly prevalent	Bucket 6: Chances of high prevalence	Bucket 9: no chance for prevalence

Information in buckets 1 to 3 should be subjected to ex-ante scrutiny, i.e., evaluating the prevalent high-stakes information to find disinformation and misinformation, while the information in buckets 4 to 6 stays on high alert.

Calibration of Enforcement (1/2)

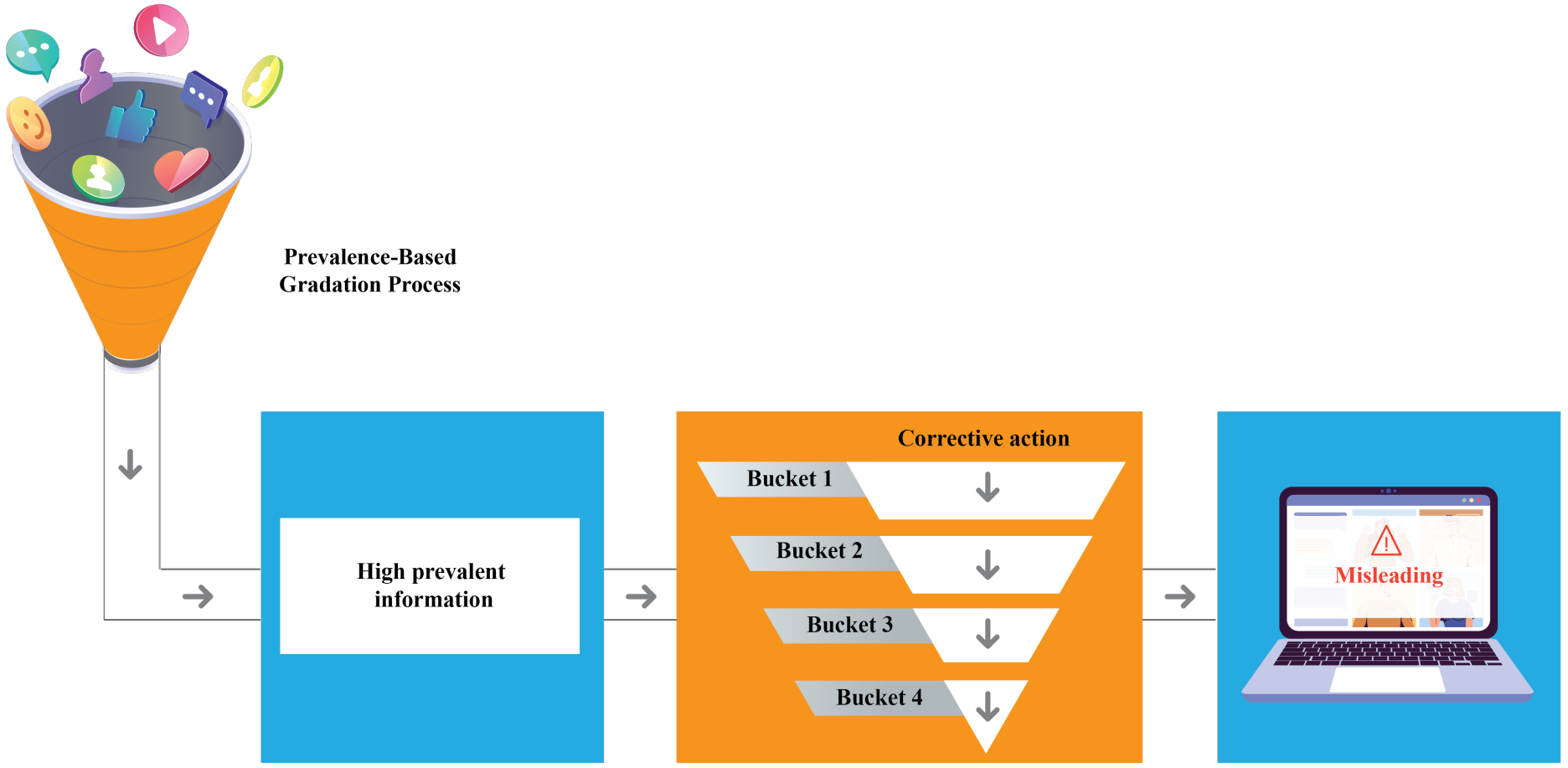


Calibration of Enforcement (2/2)

User profile / High-stake information		No. of likes, shares and comments		
		High	**Medium**	**Low**
No. of followers	**High**	Bucket 1 – Severely high prevalent	Bucket 4: Chances for severely high prevalence	Bucket 7: Fewer chances for severely high prevalence
	Medium	Bucket 2 – Very high prevalent	Bucket 5: Chances for very high prevalence	Bucket 8: fewer chances for high prevalence
	Low	Bucket 3 – highly prevalent	Bucket 6: Chances of high prevalence	Bucket 9: no chance for prevalence

Information bucket 1 should be treated differently from information in bucket 2 and subsequent buckets. For instance, bucket 3 information can start with the platform flagging and masking the information to be misleading/fake and stop people from sharing it further. As the same information starts moving to bucket 2, more severe actions can be taken.

Decision Communication



Way Forward

- Adopting a prevalence-based gradation process by social media platforms is critical. The process will help tackle false negatives, which cuts positive reinforcement for negative behaviour, i.e., posting and sharing misinformation and disinformation.
- As we move forward, the platforms must adopt this process or any other process-level intervention to strengthen the foundation that underpins the Internet's success, i.e., trustworthiness, by intervening in the application layer of the internet.
- This process-level intervention in content moderation to tackle disinformation and misinformation ensures that individuals have a trusted and secure internet experience.