India’s Deepfake Rules: Fix the Harm, Not Kill Creativity

Why blanket AI tagging backfires — and what India should do instead

What the Proposed Amendment Tries to Do

Broad definition: any algorithmically created or edited media = “synthetic”

Pre-upload duties for platforms: user self-declaration → platform verification → labeling

Toolmakers: permanent IDs + visible labels (≥10% of visual/first 10% of audio)

Safe-harbour tied to proactive removals → incentive to over-moderate

The Real Issues

Almost everything digital becomes “synthetic” → over-inclusion

Platforms can’t judge intent or authenticity at upload time

Safer to over-remove/over-label than risk liability → **chilling effect**

Still weak on holding malicious deepfake creators accountable

Why Blanket Tagging Backfires

False positives label benign AI art, thumbnails, explainers

Warning fatigue: labels get ignored; bad actors strip metadata

Creators self-censor; educational and creative uses suffer

Political risk: ‘proactive’ takedowns can skew debate in elections

No Country Tags All AI Content

USA: harm-specific, creator-liable (elections, deepfake porn, fraud) — no universal tagging

EU (DSA/AI Act): disclosures for realistic deepfakes in factual contexts; strong appeals & transparency

UK (Online Safety): targets harmful synthetic media; no blanket pre-upload tagging

Global principle: regulate deception & harm, not the mere use of AI

What India Should Aim For

Pre-emptive containment, not pre-publication censorship

Accountability on malicious creators; platforms protected for fair process

Preserve legitimate AI creativity, satire, parody and education

Middle Path: Pre-emptive Containment (Implementable)

Uploader self-declaration for realistic person/event depictions in factual contexts
Attach/verify provenance (C2PA) — metadata by default; no big overlays
Risk classify; monitor virality; apply circuit-breakers (pause recs, context banner)
Rapid human review (30–120 min): benign → restore; risky → small visible label; illegal → remove
Election guardrails: dual review, parity checks, public ledgers; hash-sharing for confirmed malicious deepfakes

Accountability, Due Process, Transparency

New offence: malicious synthetic impersonation (elections, fraud, sexual exploitation, targeted defamation)
Primary liability on creators; platform liability only for process failure
Notice → label/downrank while reviewing → reasoned decision → fast appeal (48–72h)
Monthly public stats; audited trusted-flaggers; penalties for abusive flags

Call to Redraft — A Precise, Harm-Based Law

Narrow scope: realistic emulations likely to deceive a reasonable viewer in factual contexts

Calibrated disclosures: metadata by default; visible labels only where risk is real

Process-based safe-harbour: immunity for fair workflows, not mass removals

Protect citizens from deepfakes — and protect creativity from over-regulation