

ObscureIQ Analysis: How Can We Stop DeepFakes?

Intervention Type	Intervention Point		Countermeasure Type	Countermeasure Example	Adoption Barriers	Side Effects	Pro/Reactive
Creation/Origin Ctrl	Restrict Access to Tools	🏢	Policy & Governance Solutions	Limit model releases, gate APIs	Impossible to fully control open-source	Stifles research, drives underground use	Proactive
	Restrict Training Data	🏢	Policy & Governance Solutions	Lock down facial/voice datasets	Hard to enforce globally	Limits legitimate AI development	Proactive
	Model Guardrails	🏢	Policy & Governance Solutions	Block harmful prompts in AI	Can be bypassed	Over-blocks safe use cases	Proactive
	Watermarking	⚙️	Technological Solutions	Hidden digital signatures	Needs global standardization	Can be stripped, fragility	Proactive
	Identity Verification for AI Agents	🏢	Policy & Governance Solutions	Require AI to self-disclose generated media	Open-source outcry, hard to enforce globally	Over-blocks anonymity, limits creative freedom	Proactive
	Metadata Binding (C2PA)	⚙️	Technological Solutions	Provenance + chain-of-custody	Requires broad adoption	Breaks if re-encoded, partial coverage	Proactive
	Content Provenance	⚙️	Technological Solutions	Blockchain for Verification	Scalability, high energy cost, reqs trusted network	Slow processing speeds, potential for centralization	Proactive
	Econ Sanctions Against Infra	🏢	Policy & Governance Solutions	Target hosting providers/model hubs who enable	Global enforcement issues, geopolitics	Collateral damage to benign platforms, censorship	Proactive
	Industry Standards	🏢	Policy & Governance Solutions	Watermark + provenance coalition	Requires coordination	Slow adoption	Proactive
	Government Regulation	🏢	Policy & Governance Solutions	Disclosure laws, bans	Politicized, fragmented globally	Overreach, stifled innovation	Proactive
	Device-Level Signing	🛡️	Technological Solutions	Phones/cameras cryptographically sign media	Device maker cooperation, hardware upgrades	Device costs, + digital divide, surveillance	Proactive
	Red-Team Testing Ecosystem	🛡️	Technological Solutions	Organized adversarial testing of detection models	Costly, collab across competitors	Could be weaponized, used as theater	Proactive
Platform / Dissemination Control	Trusted Media Registries	🏢	Policy & Governance Solutions	News/Gov/NGO registry of verified media	Requires trust, coordination across institutions	May entrench gatekeepers, potential bias concerns	Proactive
	Platform Upload Scanning	⚙️	Technological Solutions	AI deepfake detectors	Imperfect detection, latency	False positives, censorship concerns	Reactive
	Tiered Trust Systems	🏢	Policy & Governance Solutions	Verified vs anonymous posting	Privacy tradeoffs, user pushback	Endangers activists, reduces anonymity	Proactive
	Platform Labeling	🏢	Policy & Governance Solutions	This may be manipulated badges	User distrust of platforms	Labels ignored, polarization	Reactive
	Virality Controls	⚙️	Technological Solutions	Share limits, dampening, circuit breakers	Platforms fear engagement drop, politics	Suppresses legit viral content, “shadow ban” accusations	Reactive
	Detection Technology	🛡️	Technological Solutions	Liveness Detection and Biometrics	Privacy, sophisticated attacks can bypass	False negatives/positives, user inconvenience.	Proactive
	Detection Technology	🛡️	Technological Solutions	Adversarial Training	High compute cost, arms race w/ attackers	High dev costs, continuous updates to remain effective	Proactive
	Detection Technology	🛡️	Technological Solutions	Hybrid & Multimodal Systems	Diverse training data, high compute cost	Risk of over-general, being brittle to new vectors	Proactive
	Detection Markets/Collective Intel	🏢	Policy & Governance Solutions	Open network shared real-time detection updates	Coordination challenges, data sharing resistance	Risk of centralization, slower innovation if over-regulated	Proactive
	Rapid Takedowns	💛	Policy & Governance Solutions	Legal or platform removals	Response lag, jurisdictional limits	Whack-a-mole problem	Reactive
	Platform Transparency Portals	🏢	Policy & Governance Solutions	Public dashboards show manipulated media	Platforms fear reputational damage	Eroded trust in platforms, liability	Reactive
Reactive / Impact Mitigation	User-Facing Authenticity Signals	👤	Social & Human Solutions	Badges, provenance checks	Requires user attention & literacy	Overload, signals ignored	Reactive
	Media Literacy	📖	Social & Human Solutions	Public awareness campaigns	Slow cultural adoption	Minimal impact in real-time crises	Proactive
	Crowdsourced Verification	👥	Social & Human Solutions	Fact-check communities	Susceptible to brigading	Polarization, low trust	Reactive
	Incident Response Playbooks	🏢	Policy & Governance Solutions	Rapid-response protocols for deepfake crises	Needs cross-sector buy-in, slow policy cycles	Overreach risk, political misuse in emergencies	Reactive
	Legal Liability	🏢	Policy & Governance Solutions	Criminal + civil penalties	Enforcement across borders	Risks chilling free expression	Reactive
	Legal & Judicial Frameworks	🏢	Policy & Governance Solutions	Evolving Rules of Evidence	Slow adapt, fragged across juris, enforcement	Risk of stifling free speech, can increase litigation costs	Reactive
	Victim Support Systems	💛	Social & Human Solutions	Content suppression, PR repair	Expensive, limited access	Helps only privileged victims	Reactive
	Insurance & Risk Management	🏢	Policy & Governance Solutions	Usually org-level coverage	Reach limited to orgs and very wealthy	Treats symptom, not cause	Reactive
	Truth Anchoring	💛	Social & Human Solutions	Publish verified authentic content as baseline	Reqs consistency, credibility. Can be expensive.	Pressure for authenticity. Vulnerable to spoofing.	Both
	Pre-Bunking Campaigns	📖	Social & Human Solutions	Adv warning campaigns of expected narratives	Requires public trust in institutions	Alert fatigue, weaponization of process	Proactive