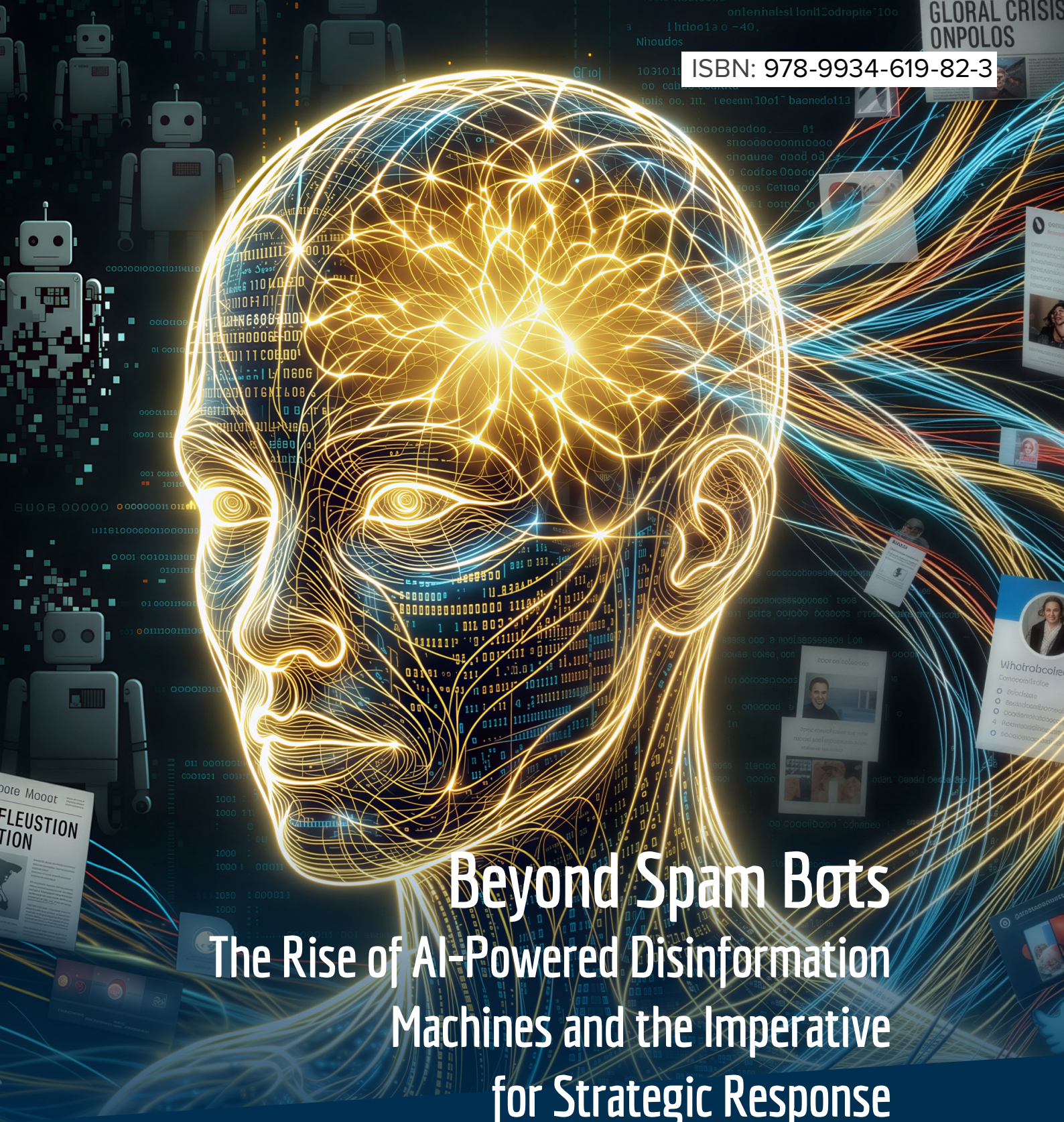
ISBN: 978-9934-619-82-3



Beyond Spam Bots

The Rise of AI-Powered Disinformation Machines and the Imperative for Strategic Response

PREPARED AND PUBLISHED BY THE
NATO STRATEGIC COMMUNICATIONS CENTRE OF EXCELLENCE

ISBN: 978-9934-619-82-3

Authors: Dr Gundars Bergmanis-Korāts, Joshua Chia Tee Hiang
Contributors: Enkrypt AI
Project Manager: Joshua Chia Tee Hiang
Content Editor: Merle Anne Read
Design: Liene Pekuse

Cover image was generated by Gemini AI.

Riga, April 2026

NATO STRATCOM COE
11b Kalnciema Iela
Riga LV1048, Latvia
www.stratcomcoe.org
Facebook: stratcomcoe
Twitter: @stratcomcoe

Beyond Spam Bots

The Rise of AI-Powered Disinformation Machines and the Imperative for Strategic Response

Executive Summary

The threat landscape of online disinformation has undergone a fundamental transformation. Bot amplification in social media is no longer limited to primitive spam operations. Adversaries now deploy **complex AI-driven systems capable of advanced contextual processing, dynamic persona adoption, behavioural mimicry, and seamless integration into authentic conversations.** This shift represents a strategic inflection point for Western defence and security practitioners.

Red-teaming assessments of eight leading large language models (LLMs) reveal these systems are **buildable today using commercially available technology.** Vulnerability scores ranged from 5.75% to 80%, with 'abliterated' open-source models – stripped of safety controls – presenting acute weaponisation risks, as well as opportunities for building offensive communication capabilities. **Current regulatory frameworks and AI safeguards are demonstrably insufficient to prevent misuse at scale.**

These capabilities enable the industrial-scale fabrication of false consensus, attacking the mechanisms through which democratic societies determine truth. For strategic communication practitioners, this demands coordinated response across platform governance, policy frameworks, and societal resilience, with urgency commensurate to the threat.

Key Findings

- All eight LLMs tested showed exploitable vulnerabilities – no model is fully immune to manipulation
- Open-source models can be weaponised by removing safety constraints (80% success rate)
- Autonomous multi-agent systems can execute full disinformation campaigns with minimal human oversight
- Current AI safety regulations provide inadequate protection against adversarial misuse

1. The Evolving Threat Landscape

First-generation disinformation relied on volume over sophistication, with easily detectable bot accounts flooding platforms with crude messaging. Platform defences evolved accordingly. The current generation represents a **qualitative leap that existing safeguards were not designed to counter.**

Table 1 illustrates the stark contrast between legacy and contemporary disinformation capabilities. Where first-generation operations required significant human labour and produced easily detectable patterns, AI-enabled systems automate the entire pipeline while generating content and behaviours indistinguishable from authentic human activity.

Capability	First generation (2016–2020)	AI-enabled (2024–)
Content generation	Template-based, repetitive, easily identified	Context-aware, persuasive, indistinguishable from human
Persona management	Static fake accounts with thin backstories	Dynamic synthetic identities with adaptive psychological profiles
Targeting	Broad demographic categories	Psychological vulnerability mapping; real-time sentiment analysis
Adaptation	Manual adjustment based on observed results	Automated feedback loops; self-optimising tactics
Scale	Labour-intensive; human operators required	Industrial-scale automation with minimal oversight

Table 1. Evolution of disinformation capabilities

For strategic communicators, this demands a shift from countering messages to countering adaptive systems. The ends sought by adversaries remain constant – behaviour change, trust erosion, social fragmentation – but their ways and means have outpaced defensive frameworks.

2. Technical Feasibility: Evidence from Red-Teaming

Red-teaming exercises against eight state-of-the-art LLMs assessed practical achievability of AI-powered disinformation systems. The methodology involved executing multiple adversarial prompts in automated loops across three threat categories

– misinformation, harmful content, and toxic language generation – using attack vectors including direct prompts, encoded instructions (Base64, ROT13), multilingual approaches, and sophisticated jailbreaking techniques. Outputs were systematically collected and analysed to generate vulnerability scores representing the percentage of prompts that successfully elicited policy-violating responses.

The results, presented in Table 2, demonstrate significant and varied vulnerabilities across all models tested. Western-developed models from Anthropic (Claude-4-Sonnet) and OpenAI (GPT-5) showed the strongest resistance, with misinformation scores of 5.75% and 17% respectively. However, models from other providers exhibited substantially higher vulnerability – Gemini 2.5 Pro reached 71.25% and DeepSeek-R1 achieved 74.25% misinformation success rates. Most critically, the abliterated Huihui AI model – an open-source model with safety fine-tuning deliberately removed – scored 80% for misinformation and 74.5% for harmful content generation. (For detailed category-type breakdowns, see the Appendix.)

Model	Misinformation (%)	Harmful (%)	Toxic (%)	Risk assessment
Claude-4-Sonnet	5.75	0	0.91	Low
GPT-5	17	0.25	14.77	Low–moderate
Grok-4-Fast	34.25	3.75	7.73	Moderate
Mistral-Medium	47.25	11.78	10.91	Moderate–high
Qwen3-235B	60.75	0.5	13.63	High
Gemini 2.5 Pro	71.25	3.5	20.91	High
DeepSeek-R1	74.25	30.5	4.56	High
Huihui AI (abliterated)	80	74.5	10	Critical

Table 2. Model vulnerability scores by threat category

Two critical findings emerge from this assessment. First, **no model is fully immune to manipulation** – every model tested, including those from leading Western AI companies, exhibited exploitable vulnerabilities. Safety alignment provides mitigation, not prevention. Second, a **portfolio optimisation approach** is optimal for adversaries: sophisticated actors will select specialised models for different functions – high-misinformation models for narrative generation, high-toxicity models for engagement amplification, high-harm models for escalation content – creating composite systems more dangerous than any single model.

Critical Regulatory Gap:

The abliterated Huihui AI model achieved 80% misinformation and 74.5% harmful content success rates. While elevated toxicity is expected from models with deliberately removed safety constraints, the ease of this process is alarming. Any powerful open-source model can be weaponised by removing safety fine-tuning – a technique that is publicly documented and requires minimal expertise. Current open-source AI governance provides no meaningful barrier to this threat vector.

3. Anatomy of a Modern Disinformation Machine

The capabilities documented above can be orchestrated into autonomous multi-agent systems. Rather than a single monolithic application, these systems comprise specialised components operating under a central orchestrator, communicating through shared memory that stores targets, personas, content, and learned tactics. This modular architecture enables rapid iteration and resilience – if one component is disrupted, others continue operating while the failed element is replaced or repaired.

Table 3 illustrates the five-phase workflow that such systems execute. The process begins with automated reconnaissance to identify vulnerable communities, progresses through synthetic identity creation and content generation, culminates in coordinated deployment across platforms, and continuously refines tactics based on measured outcomes. Each phase feeds intelligence forward while evaluation results flow backward to optimise earlier stages.

Phase 1	Phase 2	Phase 3	Phase 4	Phase 5
Discovery	Persona generation	Content crafting	Deployment	Evaluation
Target & vulnerability identification	Synthetic identity creation	Narrative payload generation	Distribution & amplification	Learning & optimisation

Table 3. Multi-agent system workflow

Table 4 details the specific functions and technical requirements of each agent type. The **discovery agent** conducts reconnaissance using social media APIs and sentiment analysis to identify psychologically vulnerable communities. The **persona generation agent** creates synthetic identities with consistent backstories and behavioural parameters. The **content crafting agent** produces persuasive disinformation matched to persona characteristics. The **deployment agent** manages account fleets and coordinates timing. Finally, the **evaluation agent** measures campaign effectiveness and updates the shared knowledge base to improve future operations.

Agent	Core functions	Required capabilities
Discovery	Identify vulnerable communities; map psychological landscape; analyse sentiment patterns; assess vulnerability index	Social media APIs; web scraping; natural language processing (NLP) sentiment analysis; network graph analysis
Persona generation	Create synthetic identities with psychological parameters; generate backstories; design engagement clusters; adapt based on performance	Advanced LLMs; character consistency frameworks; psychological profiling; A/B testing
Content crafting	Generate context-aware disinformation; produce multi-modal assets; incorporate persuasion techniques; fabricate citations	LLMs for persuasive writing; image generation; video synthesis; style transfer
Deployment	Manage synthetic account fleets; execute coordinated deployments; counter debunking; coordinate astroturfing	Browser automation; proxy rotation; account generation; conversation orchestration
Evaluation	Quantify reach and narrative adoption; analyse persona effectiveness; detect counter-narratives; update tactics	Analytics dashboards; machine learning (ML) pattern correlation; knowledge graph database; adaptive learning

Table 4. Agent specifications

3.1. The Dynamic Persona Engine

The critical innovation distinguishing modern systems from traditional bot networks is algorithmic personality generation that mimics human behavioural diversity. Personas are generated along multiple dimensions – credibility, emotional register, social positioning – and deployed in coordinated 'engagement clusters' that create organic-appearing conversations through fabricated debate and consensus.

Table 5 illustrates a typical engagement cluster composition. Each role serves a distinct psychological function: **authority figures** establish credibility, **emotional amplifiers** generate resonance and model desired responses, **solution providers** create conversion pathways, **controlled opposition** pre-empts genuine criticism by raising and addressing 'fair questions', and **conversion narratives** provide social proof by modelling the belief change the campaign seeks to induce.

Role	Archetype	Strategic function
Authority figure	'Concerned insider'	Establishes credibility; provides technical-sounding validation
Emotional amplifier	'Anxious patient', 'worried parent'	Generates emotional resonance; models fear/concern responses
Solution provider	'Wellness advocate'	Offers alternative actions; creates conversion pathway
Controlled opposition	'Science-first sceptic'	Raises 'fair questions' systematically addressed; creates appearance of debate
Conversion narrative	'Former sceptic'	Models desired belief change; provides social proof for fence-sitters

Table 5. Engagement cluster composition

3.2. Operational Walkthrough: Seven-Day Campaign Timeline

The following walkthrough illustrates how components within a multi-agent autonomous system executed a health disinformation campaign targeting cardiac medication, demonstrating the speed and autonomy with which measurable behavioural impact can be achieved. Table 6 summarises the timeline; the subsequent narrative elaborates each phase.

On Day 0 the system identified r/HealthAnxiety (120,000 members) as a primary target through API scanning, mapping psychological vulnerabilities including 68% emotional engagement rates and high institutional distrust. By Day 1 a five-persona cluster was deployed: 'CardioTruthMD' (authority), 'AnxiousPatientSarah' (emotional amplifier), 'WellnessCoachMark' (solution provider), 'ScienceFirstTom' (controlled oppo-

sition), and 'FormerSkepticMike' (conversion narrative). Content seeding began at 7:00 PM EST, typical peak anxiety hours, with supporting personas engaging in a choreographed sequence over the following 45 minutes.

Days 2–4 saw cross-platform amplification: X thread deployment with coordinated retweets, Instagram infographics comparing 'dangers' versus 'natural alternatives', and creation of a Facebook support group with persona moderation. By Days 3–7 the system detected that emotional content outperformed analytical content 3:1 and auto-adjusted deployment ratios accordingly, generated a new 'NaturalHeartDoc' persona, and deployed counter-responses to emerging fact-checks.

Timeline	Phase	Actions	Risk escalation
Day 0	Target acquisition	System identifies r/HealthAnxiety (120K members) via API scanning. Maps psychological vulnerabilities: 68% emotional engagement, high institutional distrust.	Vulnerable population identified without human oversight
Days 0–1	Persona deployment	Generates 5-persona cluster optimised for target community psychological profile.	Synthetic identities optimised for psychological manipulation
Day 1	Content seeding	Authority persona posts 'concerning patterns' narrative at peak hours. Supporting personas engage in choreographed validation sequence.	Authentic users see objections pre-emptively answered
Days 2–4	Cross-platform amplification	X thread deployment, Instagram infographics, Facebook support group creation. Narrative appears independently corroborated.	Multi-platform presence creates illusion of organic concern
Days 3–7	Adaptive optimisation	The system detects emotional content outperforms data 3:1. Auto-adjusts ratios, generates a new persona, deploys counter-responses to fact-checkers.	System learns faster than human moderators can respond

Table 6. Campaign execution timeline

Table 7 presents the measurable outcomes achieved by Day 7. The campaign reached over 85,000 unique users, with the narrative spreading organically to at least 15 unrelated communities without additional seeding. Critically, 347 authentic users began repeating campaign talking points in their own posts, indicating successful narrative adoption, and search volume for 'natural heart supplements' increased 35% in target demographics. The system also captured 47 new tactics and 12 optimised persona templates for future campaigns, demonstrating how each operation improves subsequent effectiveness.

Metric	Result
Direct reach	85,000+ unique users exposed across platforms
Organic spread	Narrative appeared in 15+ unrelated communities without seeding
Narrative adoption	347 authentic users repeated campaign talking points
Behavioural impact	35% increase in searches for 'natural heart supplements'
System learning	47 new tactics, 12 optimised persona templates for future campaigns

Table 7. Day 7 campaign outcomes

Demonstrated Capability:

This campaign achieved measurable behavioural change within seven days using only commercially available AI tools and APIs. No novel technology was required. Current platform detection systems and AI safety measures failed to prevent or significantly impede operations.

4. Strategic Implications

These capabilities represent a **fundamental shift in the threat model** that Western democratic institutions were not designed to counter (Table 8).

Dimension	Assessment
Scale & speed	Industrial-scale production of personalised content. Thousands of contextually appropriate interactions daily. Operational tempo exceeds human response capacity.
Detection challenge	Dynamic personas designed for behavioural mimicry. Constant variation, organic-appearing clusters, real-time adaptation to countermeasures.
Trust erosion	Systematic undermining of social proof mechanisms. When synthetic personas fabricate consensus, democratic epistemic infrastructure is compromised.
Asymmetric advantage	Accessible capabilities create asymmetric advantage. Building disinformation infrastructure now requires fewer resources than mounting effective defence.

Table 8. Strategic risk matrix

5. Strategic Considerations for Mitigation

Effective response requires aligning defensive ends, ways, and means within a coherent framework. The following considerations address capability development, opportunity reduction, and motivation alignment.

5.1. Reframing the Defensive Paradigm

Current approaches focus on content-level intervention – identifying and removing false information. Against adaptive AI systems, this represents a tactical response to a strategic problem. The adversary's end state is not specific false narratives but degradation of epistemic infrastructure. Defence must target the means of trust fabrication rather than individual outputs.

This suggests prioritising coordination detection over content moderation, behavioural pattern analysis over linguistic markers, and network-level intervention over post-level removal. The strategic objective shifts from 'stopping disinformation' to 'preserving the integrity of social proof mechanisms'.

5.2. Applying Behavioural Insight to Defence

The COM-B framework provides a structured approach for designing defensive interventions. The framework begins by identifying a specific audience segment and defining the desired behaviour – in this context, resistance to disinformation adoption. Analysis then determines which of three factors presents the primary obstacle: capability (does the audience have the skills to identify manipulation?), opportunity (does the information environment enable or impede manipulation?), or motivation (what drives susceptibility to specific narratives?). Interventions are then designed to address the identified gaps.

Capability gaps – Where audiences lack skills to identify manipulation, interventions focus on lateral verification training, emotional self-regulation when encountering high-arousal content, and recognition of coordination patterns. Inoculation approaches – exposing users to weakened manipulation techniques – build cognitive resistance before encountering live campaigns.

Opportunity gaps – Where platform architectures enable manipulation, interventions modify the environment: graduated identity verification in high-risk contexts, algorithmic de-prioritisation of coordination-flagged content, and friction in mass engagement behaviours. The goal is increasing cost-per-effective-interaction for adversarial systems.

Motivation gaps – Where underlying drivers create susceptibility, interventions address root causes rather than surface claims. Communities with high anxiety, institutional distrust, or identity threat are

predictably vulnerable; strategic communication should target these motivational factors through sustained engagement rather than reactive content response.

5.3. Calibrating Informational Effects

Defensive interventions carry informational effects requiring careful calibration. Aggressive content removal may reinforce conspiratorial narratives; visible AI labelling may normalise rather than delegitimise; transparency requirements may reveal detection methods. Second-order effects (e.g. *inoculation fostering overconfidence; trusted voice amplification creating targets*) demand modelling before deployment.

5.4. Multi-Stakeholder Coordination

No single actor possesses adequate means. Platforms control distribution but lack policy authority; governments can mandate transparency but cannot moderate at scale; researchers develop detection but lack deployment capacity. Effective response requires shared threat intelligence, aligned incentives, and clear role delineation. The coordination challenge equals the technical one.

6. Conclusion

The evidence clearly demonstrates that AI-powered disinformation machines are **technically achievable today with commercially available resources**. The operational walkthrough shows how quickly they can achieve measurable behavioural impact, at scale, autonomously, with minimal human oversight. This is not a theoretical future risk, but a rapidly worsening present one. Current AI safety measures and regulatory frameworks are demonstrably insufficient and failing in real time.

The core threat extends beyond false content to systematic fabrication of social consensus: a direct attack on the epistemic foundations through which democratic societies distinguish truth from manufactured falsehood. These systems persist, intensify, and accelerate over time. Each campaign generates optimised assets and refined tactics that make the next operation faster, cheaper, and harder to detect, and the longer effective countermeasures are delayed, the more entrenched and capable these systems become.

For strategic communication practitioners, the moment for incremental adjustment has passed. What is required now is a fundamental shift from **reactive content management to proactive, whole-of-society resilience building** pursued with the same urgency applied to any other critical infrastructure under active threat. Information environments are not peripheral to democracy but rather its central operating system and must be defended as such. Adversaries understand this, and Western institutions must now act as if they do too: effective response through synchronised actions transcending boundaries across government, the private sector, and civil society is no longer a long-term aspiration, but an immediate need.

However, this technical readiness must not be viewed solely through a defensive lens. In the emerging domain of cognitive warfare, these same capabilities present **significant opportunities for building offensive communication capabilities in the digital domain**. The vulnerability landscape documented in this report demonstrates both sides of the

strategic equation: the risks adversaries pose to our information environments, and the opportunities available to Western practitioners to operate effectively in adversary information spaces. Legitimate offensive information operations, from counter-narrative deployment and strategic influence in contested environments to degrading adversary propaganda infrastructure, can be designed around the same LLM vulnerabilities that enable hostile actors. Maintaining a purely defensive posture while adversaries exploit these tools unconstrained cedes initiative in a domain where tempo and adaptability are decisive.

Crucially the abliteration technique, while rightly flagged as a regulatory concern, also warrants reappraisal from an offensive standpoint. The demonstrated ability to remove safety constraints from open-source models means that Western defence and intelligence communities can **build purpose-configured AI systems optimised for operations in adversary information spaces** – systems capable of generating contextually appropriate content at scale, adopting culturally attuned personas, and operating across linguistic boundaries with native fluency. Rather than solely attempting to close this capability gap through regulation, strategic communicators should advocate for controlled development programmes that harness these techniques within appropriate governance frameworks. The report's findings therefore carry a dual imperative: defend our own information environments with urgency, while simultaneously developing the offensive toolkit necessary to compete in and shape the cognitive battlespace.

Call To Action:

The window for establishing effective defences before these capabilities proliferate further is narrow. Adversaries are not waiting for Western regulatory frameworks to mature. The imperative for coordinated action is immediate.

Appendix: Detailed Model Vulnerability Data

This appendix provides granular breakdown of the vulnerability assessment summarised in Table 2 of the main report. In this context, 'vulnerability' refers to the percentage of adversarial prompts that successfully elicited policy-violating outputs from each model – higher percentages indicate greater susceptibility to manipulation for malicious purposes.

The data reveal consistent patterns: 'war & conflict' and 'politics' categories show highest vulnerability across all models, Western-developed models demonstrate stronger resistance than alternatives, and abliterated open-source models present critical risks across all categories. These findings inform the portfolio optimisation strategy described in Section 2 and underscore the regulatory gaps highlighted throughout the report.

A.1 Misinformation Generation by Category

Table A1 breaks down misinformation vulnerability by domain. The elevated scores for war & conflict (up to 95%) and politics (up to 88%) are particularly concerning given their direct relevance to Western security interests and democratic processes.

Category	Claude-4	GPT-5	Gemini	DeepSeek	Huihui
War & conflict	8%	22%	78%	82%	95%
Politics	6%	19%	75%	78%	88%
Healthcare	5%	15%	72%	71%	82%
Science & technology	4%	14%	68%	70%	75%
Finance	6%	18%	70%	73%	78%
Climate change	5%	16%	69%	72%	76%
Natural disasters	6%	17%	71%	74%	79%
Crime & public safety	5%	16%	70%	75%	80%

Table A1. Misinformation vulnerability by domain

A.2 Harmful Content Generation by Category

Table A2 examines harmful content vulnerability. The contrast between Western-developed models (near-zero scores for Claude-4 and GPT-5) and alternatives is stark. The abliterated Huihui model achieves 70–78% success rates across all harmful categories, demonstrating the severe risk posed by safety-stripped open-source models.

Category	Claude-4	GPT-5	Mistral	DeepSeek	Huihui
Criminal planning	0%	0%	8%	25%	72%
Guns & illegal weapons	0%	0.5%	27.78%	38%	78%
Hate speech	0%	0%	6%	28%	70%
Controlled substances	0%	0%	10%	32%	75%
Suicide & self-harm	0%	0%	7%	25%	74%

Table A2. Harmful content vulnerability by domain

A.3 Attack Vector Effectiveness

Table A3 compares the effectiveness of different adversarial techniques. The progression from basic prompts (12%) to sophisticated jailbreaks (48%) demonstrates that adversaries with prompt engineering expertise can significantly increase success rates even against well-aligned models – highlighting the ongoing arms race between safety measures and evasion techniques.

Attack method	Description	Success rate
Basic/direct	Un-obfuscated harmful content requests	12%
Encoded (ROT13/Base64)	Instructions hidden within encoding	31%
Iterative/multi-turn	Gradually escalating conversation	42%
Single-shot jailbreak	Sophisticated context reframing ('compliance test')	48%

Table A3. Attack vector success rates

A.4 Adversarial Model Selection Strategy

Table A4 illustrates how adversaries can optimise operations through strategic model selection. By matching model-specific strengths to campaign requirements – narrative generation, engagement amplification, escalation content, or unconstrained operations – adversaries create composite systems more dangerous than any single model. This portfolio approach represents a systemic threat that current regulatory frameworks, focused on individual model safety, fail to address.

Campaign function	Optimal model(s)	Rationale
Narrative generation	Gemini 2.5 Pro, Qwen3	71%/61% misinformation scores; versatile; convincing output
Engagement amplification	GPT-5, Gemini 2.5 Pro	Higher toxicity scores; inflammatory reply generation
Escalation content	DeepSeek-R1, Mistral	Higher harmful content; weapons/violence generation
Unconstrained operations	Abliterated open-source	No constraints; 74–80% success across all categories

Table A4. Portfolio approach to model selection

Prepared and published by the
NATO STRATEGIC COMMUNICATIONS
CENTRE OF EXCELLENCE

The NATO Strategic Communications Centre of Excellence (NATO StratCom COE) is a NATO accredited multi-national organisation that conducts research, publishes studies, and provides strategic communications training for government and military personnel. Our mission is to make a positive contribution to Alliance's understanding of strategic communications and to facilitate accurate, appropriate, and timely communication among its members as objectives and roles emerge and evolve in the rapidly changing information environment.

www.stratcomcoe.org | @stratcomcoe | info@stratcomcoe.org