International Journal of Education, Modern Management, Applied Science & Social Science (IJEMMASSS)
ISSN :2581-9925, Impact Factor (Cosmos: 7.555 & I2OR: 5.105), Volume 08, No. 01(III) (Special Issue), January- March, 2026, pp. 23-34

National Security in a Borderless Web: Examining the Interplay Between Social Media Tools and Enhanced Digital Espionage

Dr Ashok Kumar[1] | Ms. Usha[2*] | Dr. Neelu Grover[3]

[1]Director, Geopolitics & Defence study Research Cell, JNV University, Jodhpur (Raj.).
[2]Student, Geopolitics & Defence study Research Cell, JNV University, Jodhpur (Raj.).
[3]Associate Professor, Motilal Nehru College (Eve), University o Delhi.

[*]Corresponding Author: ushaarjunchoudhay@gmail.com

Abstract

The rapid evolution of social media has transformed the digital landscape into a "borderless web," dissolving traditional geopolitical frontiers and creating a porous environment where state and non-state actors engage in unprecedented levels of mutual surveillance. This paper investigates the dual-use nature of social platforms—tools designed for connectivity that have been weaponized into sophisticated engines for digital espionage. We analyze how emerging technologies, specifically Generative AI (GenAI), Deepfake synthesis, and automated reconnaissance, have lowered the barrier to entry for espionage, enabling high-fidelity social engineering at a machine-driven scale. Reactive security models are increasingly obsolete against these "machine-speed" threats, which exploit the inherent trust and data surplus of social ecosystems. To mitigate these vulnerabilities, this paper proposes the Real-Time Adaptive Intelligence &Defence (RAID) Framework. Unlike static defences, RAID integrates:

- *AI-Powered Behavioural Analytics: To detect non-linear patterns of intrusion and social engineering in real time.*
- *Predictive Game-Theoretic Modeling: To anticipate attacker trajectories and dynamically reconfigure security protocols.*
- *Automated Micro-segmentation: To isolate compromised data streams instantly upon threat detection.*

The study concludes that as the web facilitates borderless interaction, national security must shift from a "perimeter-defence" mindset to a "continuous-verification" model. The RAID framework provides a proactive architecture capable of safeguarding national and individual integrity in an era of transparent borders.

***Keywords**: Digital Espionage, Social Media Intelligence, RAID Framework, Generative AI, Deepfakes, National Security, Cybersecurity, Social Engineering, OSINT, Micro-segmentation, Game-Theoretic Modeling.*

Introduction

The contemporary geopolitical environment is defined by an increasingly paradoxical condition: as digital connectivity intensifies global interdependence, it simultaneously amplifies vulnerabilities that nation-states, corporations, and individuals cannot adequately defend against using conventional security architectures. The proliferation of social media platforms—from established giants such as Facebook, X (formerly Twitter), LinkedIn, and Instagram to emerging encrypted ecosystems like Telegram and Signal—has fostered a "borderless web" in which information flows freely, identities are fluid, and geographic boundaries are rendered operationally irrelevant (Rid, 2020).

[*]***A Special Issue for the papers accepted for publication in the International Multidisciplinary Conference on India's Foreign Policy and International Relations:A Future Roadmap (IFPIR-2026) (Hybrid Mode)(22-24 January 2026) Organized by Geo-Politics & Defence Study Research Cell, Faculty of Arts, Education and Social Sciences, Jai Narain Vyas University, Jodhpur, Rajasthan.***

This borderless quality, while democratising access to information and enabling grassroots political participation, has concurrently become the most significant force multiplier in the contemporary espionage landscape. Intelligence agencies and hostile non-state actors alike have recognised that the vast reservoirs of openly available personal, organisational, and behavioural data on social platforms constitute an intelligence goldmine—accessible without the risks traditionally associated with human intelligence (HUMINT) operations. The shift from embassy-based spying to social media-enabled Open-Source Intelligence (OSINT) represents a fundamental transformation in the grammar of espionage (Taddeo, 2017).

The acceleration of this transformation is directly attributable to three technological vectors: Generative Artificial Intelligence (GenAI), which enables synthetic persona creation and automated engagement at previously impossible scales; Deepfake synthesis technology, which can fabricate compelling audiovisual evidence of statements never made; and automated reconnaissance tools, which can map entire social networks, identify high-value targets, and extract sensitive data through cross-platform correlation. Together, these technologies have not merely lowered the cost of espionage—they have industrialised it.

The national security implications of this industrialisation are profound. State actors—most notably those aligned with Russian, Chinese, Iranian, and North Korean intelligence services—have demonstrated sophisticated operational use of social platforms for influence operations, credential harvesting, and strategic intelligence collection (Nimmo, 2022). However, the democratisation of these tools means that well-funded non-state actors, criminal networks, and even ideologically motivated individuals can now conduct operations that were previously the exclusive province of nation-states.

This paper makes three primary contributions to the field. First, it offers a systematic analysis of the mechanisms by which social media platforms have been weaponised for digital espionage, with particular attention to the technological amplifiers that have transformed the threat landscape between 2018 and 2025. Second, it critically examines the inadequacy of reactive, perimeter-based security models in addressing these threats. Third, it proposes the Real-Time Adaptive Intelligence & Defence (RAID) Framework—a proactive, multi-layered security architecture designed to counter machine-speed threats through continuous behavioural analysis, predictive modelling, and automated containment.

The remainder of this paper is structured as follows: Section 2 reviews relevant literature; Section 3 examines the dual-use nature of social media; Section 4 analyses the technological enablers of modern digital espionage; Section 5 presents documented case studies; Section 6 details the RAID Framework; Section 7 addresses implementation challenges; Section 8 offers policy recommendations; and Section 9 concludes.

Literature Review

The scholarly literature on the intersection of social media and national security spans multiple disciplines—intelligence studies, cybersecurity, political science, and communication theory—and has expanded rapidly since the mid-2010s as documented cases of social media-based interference became impossible to ignore.

- **Social Media as a Security Threat Vector**

Foundational work by Rid (2020) in "Active Measures" chronicles the long history of information warfare, demonstrating that social media has not invented information operations but has radically amplified their reach and precision. Similarly, Prier (2017) introduced the concept of "commanding the trend" to describe how state actors exploit social platforms' algorithmic amplification to propagate disinformation at negligible cost. These works established the conceptual groundwork for understanding social media as a dual-use infrastructure.

Chesney and Citron (2019) provided a seminal analysis of the national security implications of deepfakes, arguing that synthetic media poses a "liar's dividend"—enabling not only fabrication but also the delegitimisation of authentic evidence, with profound consequences for geopolitical stability and judicial processes. Their work has been foundational in framing the epistemological crisis created by AI-generated content.

- **OSINT and Social Engineering in the Digital Age**

Omand, Bartlett, and Miller (2012) pioneered the academic treatment of Social Media Intelligence (SOCMINT) as a distinct discipline within the intelligence community, arguing that its

collection and analysis required new ethical frameworks and legal authorities. Their work has since been extended by Macnish (2018) to address the surveillance-rights balance and by Rudner (2017) to examine state-sponsored SOCMINT operations.

The literature on social engineering has evolved substantially with the integration of AI. Hatfield (2018) identified "trust exploitation" as the primary mechanism of social engineering attacks, a finding that takes on new dimensions when trust is algorithmically simulated by GenAI systems. More recently, Mirsky and Lee (2021) provided a comprehensive taxonomy of deepfake attack vectors, classifying them by generation technique, target category, and intended harm—a framework that has been widely adopted in the cybersecurity community.

- **Limitations of Existing Frameworks**

The dominant paradigm in cybersecurity remains oriented toward perimeter defence—the identification and fortification of network boundaries to prevent unauthorised access. Schneier (2015) argued persuasively that this model is fundamentally incompatible with the distributed, always-on nature of social ecosystems, where the attack surface is effectively boundless. The Zero Trust Architecture (ZTA) model, formalised by the National Institute of Standards and Technology (NIST) in 2020, represents a significant evolution, but its implementation has been largely confined tointernal network architectures rather than extended to the open social media environment (Rose et al., 2020).

A critical gap in the existing literature is the absence of integrated frameworks that address the social media threat surface holistically—combining behavioural analytics, predictive adversarial modelling, and automated response in a single coherent architecture. Existing proposals tend to address individual components in isolation. The RAID Framework proposed in this paper is designed to fill this gap.

The Dual-Use Nature of Social Media Platforms

Social media platforms were architecturally designed to facilitate human connectivity through the sharing of personal information, interests, locations, and social relationships. This design imperative—maximising engagement through data disclosure—creates an inherent and irresolvable tension with security and privacy. The same features that make these platforms valuable for personal expression and civic participation render them structurally exploitable for intelligence collection and social engineering.

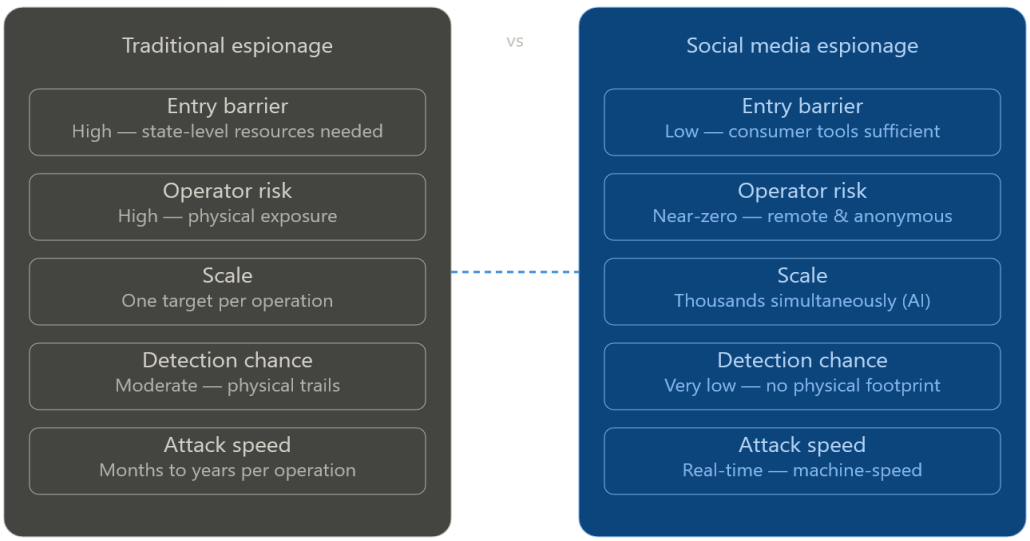

Figure 1*:* Traditional vs social media espionage comparison

- **The Data Surplus Problem**

Each user interaction on a social platform—a like, a share, a comment, a location check-in, a network connection—contributes to what we term the "data surplus": an ever-growing repository of

personally identifiable information (PII) and behavioural metadata that vastly exceeds what any individual knowingly chooses to disclose. Studies have demonstrated that a corpus of as few as 150 Facebook likes can predict a user's personality traits more accurately than their colleagues, and as few as 300 likes can outperform their spouses (Youyou, Kosinski & Stillwell, 2015).

For intelligence purposes, this data surplus enables a practice known as "mosaic theory" intelligence—the aggregation of individually innocuous data points into a comprehensive profile that reveals sensitive patterns. Military personnel who use fitness tracking applications on deployment have inadvertently mapped classified base perimeters. Defence ministry employees who disclose their LinkedIn employment history have enabled foreign intelligence services to identify key personnel for targeting. The aggregate of individually unremarkable disclosures constitutes a strategic intelligence asset.

- **Trust Architecture as an Attack Surface**

The social capital embedded in online networks—the trust relationships between followers, friends, and professional connections—constitutes an attack surface that has no parallel in traditional cybersecurity. A message delivered through a trusted social connection exploits the psychological heuristic of source credibility, dramatically increasing the probability of compliance with malicious requests. This is the foundational mechanism exploited in spear-phishing attacks, business email compromise schemes, and social engineering operations.

The platform's algorithmic architecture compounds this vulnerability. Recommendation systems designed to surface content that maximises engagement also create "filter bubbles" and polarised information environments that render populations more susceptible to confirmation-bias-driven disinformation. A target population primed by algorithmic curation to distrust official sources is significantly more receptive to synthetic disinformation tailored to their existing beliefs—a condition that intelligence services have explicitly sought to exploit.

- **Jurisdictional Gaps and Regulatory Arbitrage**

The borderless nature of social media creates fundamental challenges for national legal frameworks designed to regulate communication and intelligence collection within territorial boundaries. Content hosted in one jurisdiction, accessed in another, and created by actors in a third presents near-insurmountable jurisdictional challenges for law enforcement and regulatory agencies. State actors operating through social platforms located in permissive or uncooperative jurisdictions effectively enjoy legal immunity for operations that would constitute violations of sovereignty if conducted through physical means.

Technological Enablers of Modern Digital Espionage

The transformation of social media from a passive intelligence source to an active espionage infrastructure has been driven by three technological developments that have converged in the period since 2020. Each represents a qualitative shift rather than merely a quantitative improvement in capability.

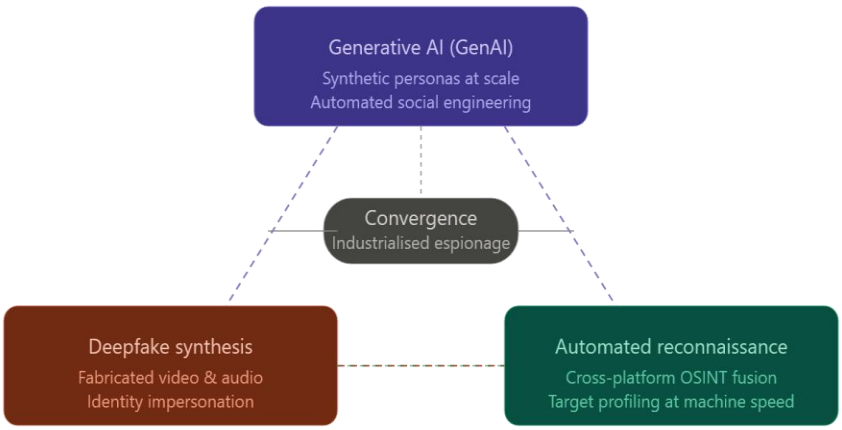

Figure 2*:* **Technology threat triad (GenAI + Deepfake + Recon)**

- **Generative Artificial Intelligence and Synthetic Persona Operations**

The emergence of large language models (LLMs) such as GPT-4, Claude, and their numerous fine-tuned variants has fundamentally altered the economics and scalability of social engineering operations. Prior to GenAI, the creation of convincing fake social media personas required sustained human effort—crafting plausible backstories, generating culturally appropriate content, managing realistic interaction patterns, and maintaining consistency across engagements. GenAI systems can now perform all of these functions autonomously and simultaneously across thousands of accounts.

The operational implications are substantial. A single human handler can now direct what effectively constitutes a battalion of synthetic social agents, each capable of conducting targeted, contextually appropriate conversations with high-value targets. The 2022 Stanford Internet Observatory report documented networks of GenAI-generated personas conducting coordinated influence operations across multiple platforms, with content quality indistinguishable from authentic human-generated material to casual inspection. More critically, GenAI enables the rapid adaptation of messaging in response to target responses—creating a feedback loop that approximates genuine relationship development.

For espionage purposes specifically, GenAI-powered personas are increasingly deployed in "honey trapping" operations—establishing romantic or professional relationships with targets to extract sensitive information or coerce cooperation. The psychological intimacy generated through sustained AI-mediated interaction can be as coercively effective as physical relationships, while eliminating the operational risks associated with human intelligence officers.

- **Deepfake Synthesis and the Weaponisation of Trust**

Deepfake technology—the use of deep learning models, specifically Generative Adversarial Networks (GANs) and diffusion models, to synthesise realistic audiovisual content—has crossed a threshold of quality that makes detection by unaided human perception unreliable. The computational cost of generating convincing deepfakes has declined from requiring specialist hardware and weeks of processing time in 2018 to being achievable on consumer-grade graphics processing units in hours as of 2024.

The national security applications of deepfake technology span a wide spectrum. At the strategic level, deepfake content purporting to show heads of state issuing inflammatory statements or military commanders issuing surrender orders could precipitate real-world crises before verification is possible. This is the "liar's dividend" identified by Chesney and Citron (2019): even after deepfakes are debunked, the cognitive residue of having seen apparently authentic evidence of an event shapes perception. At the operational level, deepfake audio is increasingly used in voice phishing (vishing) attacks that impersonate trusted authority figures—executives, commanders, or family members—to authorise transfers of funds or sensitive information.

The 2023 case in which a finance employee in Hong Kong transferred the equivalent of USD 25 million following a deepfake video conference with individuals synthesised to resemble the company's CFO and other executives represents a watershed moment—demonstrating that deepfake-facilitated fraud has matured from a theoretical concern to a financially devastating operational reality. Military and government environments face analogous vulnerabilities when authenticating sensitive communications.

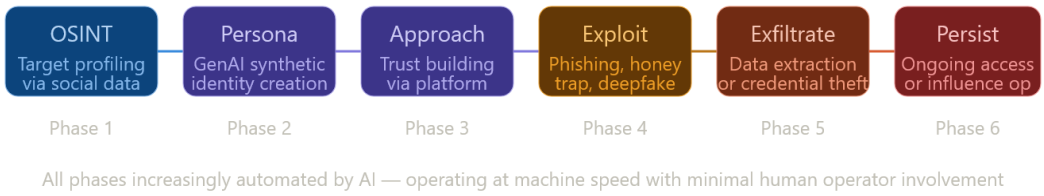

Figure 3*:* **Espionage attack chain on social media**

- **Automated Reconnaissance and Cross-Platform Intelligence Fusion**

The third technological enabler is the development of automated tools capable of conducting comprehensive reconnaissance across multiple social platforms simultaneously, fusing disparate data points into actionable intelligence profiles at machine speed. Tools ranging from commercial platforms like Maltego and Palantir to purpose-built intelligence software enable operators to map entire organisational social graphs, identify individuals with access to sensitive systems, track their physical

movements through geotagged content, identify their personal relationships and potential pressure points, and monitor their sentiment and political reliability.

The integration of AI-driven natural language processing allows these tools to analyse the content of public social media posts at scale, inferring not merely what individuals say but what they know, whom they associate with, and how they might be susceptible to manipulation. Cross-referencing public social media profiles against data breach repositories—which contain hundreds of billions of credentials from past compromises—allows automated tools to construct authentication attack packages tailored to individual targets. The combination of social context and technical credentials constitutes a highly effective attack vector against government and corporate targets alike.

Documented Case Studies of Social Media-Enabled Espionage

The theoretical vulnerabilities described in preceding sections are not speculative. A substantial body of documented evidence demonstrates that state and non-state actors have operationalised social platforms as primary espionage infrastructure. The following case studies illustrate the range, sophistication, and consequence of these operations.

Table 1*:* Case Studies Summary

Case	Year	Actor	Platform used	Method	Impact
Operation Ghostwriter	2020–23	Belarus/Russia (state)	Facebook, Twitter, news sites	Fabricated govt. documents; synthetic amplification	NATO public trust erosion; Poland, Lithuania, Latvia targeted
LinkedIn Defence Targeting	2021–22	China MSS (state)	LinkedIn	Fake headhunter/researcher personas; spear-phishing	Classified defence personnel recruited; MI5 &DGSI public warnings
AIIMS Delhi Breach	2022	Unknown (suspected state)	LinkedIn + email	Social media recon → spear-phish → ransomware	40M records compromised; hospital services disrupted weeks
India CIB Networks	2021–24	Domestic & foreign actors	Facebook, WhatsApp, Instagram	Coordinated inauthentic behaviour; algorithmic amplification	Communal tension amplified; electoral discourse manipulated
HK Deepfake CFO Fraud	2023	Criminal network	Video conferencing	Deepfake video call impersonating CFO and executives	USD 25 million transferred; watershed deepfake fraud case

- **Operation Ghostwriter: Narrative Manipulation at Scale (Belarus/Russia, 2020–2023)**

Operation Ghostwriter, first documented by Mandiant in 2020, represents a systematic, state-directed campaign to undermine NATO cohesion through the fabrication of content attributed to legitimate government and military sources. The operation combined website compromises with social media amplification to propagate fabricated quotes attributed to NATO officials, forged government documents, and synthetic news articles. The campaign specifically targeted Polish, Lithuanian, and Latvian audiences with content designed to erode confidence in Western alliance commitments.

The social media component of Ghostwriter exploited both the platforms' content distribution mechanisms and their authentication weaknesses. Fabricated content was seeded through accounts that had established credibility over years of authentic activity—demonstrating the strategic value of patience in social media influence operations. The case illustrates the asymmetric cost-benefit ratio of social media-based information warfare: the cost to the adversary of fabricating and distributing misleading content was orders of magnitude less than the cost to target nations of detecting, attributing, and countering it.

- **LinkedIn Targeting of Defence and Intelligence Personnel**

Multiple national intelligence services—including those of the United States, Germany, France, and India—have documented systematic efforts by Chinese intelligence services to recruit defence and intelligence personnel through LinkedIn. The MI5 (UK) and DGSI (France) issued formal public warnings in 2021 and 2022 respectively, identifying a pattern in which sophisticated fake profiles posing as executive headhunters, think-tank researchers, or conference organisers approached individuals with access to sensitive government systems.

The sophistication of these operations reflects the application of OSINT-driven targeting: potential recruits were identified through cross-referencing LinkedIn employment history with military or government affiliation indicators, their research interests and public statements were analysed to identify ideological or financial pressure points, and initial outreach was tailored to exploit identified vulnerabilities. Estimated recruitment successes from publicly disclosed cases suggest that these operations achieved a hit rate comparable to traditional HUMINT recruitment at a fraction of the cost and risk.

- **The AIIMS New Delhi Breach and Social Engineering Prelude (India, 2022)**

The ransomware attack on the All India Institute of Medical Sciences (AIIMS) New Delhi in November 2022—which compromised approximately 40 million patient records and disrupted critical hospital services for weeks—was preceded by a social engineering campaign that exploitedsocial media to map the institution's staff structure and identify individuals with elevated system access. Post-incident forensic analysis indicated that attackers had conducted extensive LinkedIn reconnaissance of AIIMS IT and administrative staff prior to the intrusion, enabling a targeted spear-phishing campaign that achieved initial access through a compromised staff credential.

This case is particularly instructive for the Indian national security context because it demonstrates the convergence of social media-enabled reconnaissance with critical infrastructure attack. It also highlights the inadequacy of perimeter-based defences when the attack enters through an authenticated human vector—a vulnerability that no amount of network hardening can address without complementary behavioural analytics.

- **Coordinated Inauthentic Behaviour Against Indian Democratic Processes**

Meta's Transparency Reports for 2021–2024 document multiple coordinated inauthentic behaviour (CIB) networks targeting Indian political discourse, originating from both domestic and foreign sources. These networks employed synthetic accounts to amplify polarising content related to communal tensions, electoral processes, and India-Pakistan relations, exploiting algorithmic amplification to achieve disproportionate reach relative to their actual operator numbers. The cross-cutting nature of these operations—exploiting both domestic political divisions and external geopolitical fault lines—illustrates the strategic sophistication with which social platforms are instrumentalised in hybrid warfare campaigns targeting democratic nations.

The Raid Framework: A Proactive Defence Architecture

The preceding analysis establishes that the threat posed by social media-enabled digital espionage is not amenable to solution through incremental improvements to existing reactive security models. The speed, scale, and adaptability of machine-driven adversarial operations require a qualitatively different defensive architecture—one that anticipates rather than merely reacts, that operates continuously rather than episodically, and that adapts autonomously rather than requiring constant human intervention.

The Real-Time Adaptive Intelligence & Defence (RAID) Framework is proposed as such an architecture. RAID is not a single technology but an integrated system of capabilities, each addressing a distinct dimension of the threat, designed to function synergistically as a coherent defensive whole. The framework is grounded in three core principles: continuous verification, adaptive response, and predictive posture.

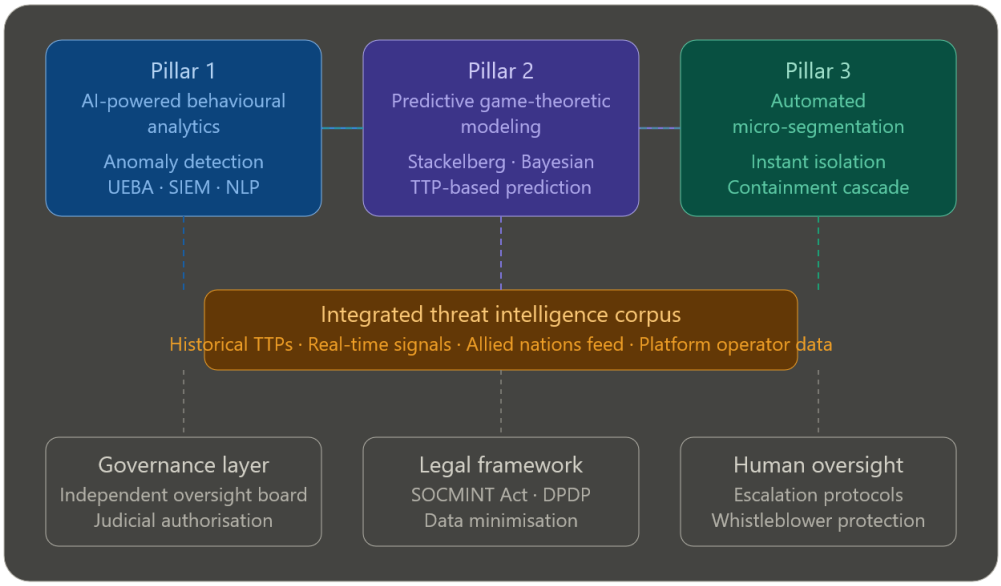

Figure 4: RAID Framework architecture

- **Pillar One: AI-Powered Behavioural Analytics**

The foundational layer of the RAID Framework is a continuous behavioural analytics capability that establishes and maintains dynamic baselines of normal activity for monitored individuals, accounts, and network segments, enabling the real-time detection of anomalous patterns that may indicate compromise or manipulation.

Unlike rule-based intrusion detection systems that flag predefined signatures of known attack patterns, AI-powered behavioural analytics employs unsupervised machine learning models—specifically autoencoders, recurrent neural networks, and transformer-based sequence models—todetect deviations from established behavioural norms that may not match any previously catalogued attack pattern. This capability is particularly critical for detecting the novel attack vectors that machine-generated espionage continuously creates.

In the social media context, behavioural analytics operates across multiple signal dimensions: communication pattern analysis (timing, frequency, network reach, sentiment trajectory), content analysis (lexical patterns, topic drift, referential anomalies), network topology analysis (connection acquisition patterns, group membership changes), and cross-platform consistency analysis (identification of synthetic personas through behavioural inconsistencies that individual platform analysis cannot detect). The fusion of these dimensions enables detection of both account compromise and synthetic persona operations with significantly higher accuracy than single-signal approaches.

A critical design requirement for this pillar is the preservation of civil liberties and privacy rights. The RAID Framework specifies that behavioural analytics should operate on aggregate behavioural signatures rather than content surveillance, and that monitoring of civilian communications should be subject to judicial authorisation and strict data minimisation principles. The framework explicitly rejects mass surveillance architectures in favour of targeted monitoring bounded by legal process.

- **Pillar Two: Predictive Game-Theoretic Modeling**

The second pillar of RAID addresses the inherent asymmetry between attacker and defender in social media espionage: attackers can observe and adapt their strategies in response to defender behaviour, while defenders who rely on reactive models are perpetually behind the curve. Game-theoretic modelling provides a formal framework for reasoning about adversarial interactions as strategic games, enabling defenders to anticipate attacker trajectories and reconfigure defences proactively.

The RAID Framework employs a multi-level game-theoretic architecture. At the strategic level, Stackelberg game models are used to optimise resource allocation across defensive capabilities,

accounting for the adversary's anticipated response to observed defensive postures. At the operational level, dynamic games with incomplete information model the evolution of ongoing influence operations and intrusion campaigns, enabling prediction of likely next moves based on observed adversary behaviour. At the tactical level, Bayesian game models assess the probability that specific observed social media accounts or content represent synthetic operations, enabling prioritised investigation and response.

The practical implementation of predictive modelling requires a substantial and continuously updated threat intelligence corpus—historical data on adversary tactics, techniques, and procedures (TTPs), combined with real-time signals from behavioural analytics, to calibrate model parameters. The RAID Framework specifies a federated intelligence architecture in which threat intelligence is shared between government agencies, platform operators, and allied nations' security services through structured protocols, while maintaining appropriate compartmentalisation of sensitive sources and methods.

- **Pillar Three: Automated Micro-Segmentation**

The third pillar addresses the consequence management dimension of social media espionage—specifically, the challenge that by the time a breach or influence operation is detected, significantdamage has typically already occurred through the propagation of compromised information or synthetic content through trusted networks. Automated micro-segmentation provides a mechanism for instantaneous containment that operates at machine speed, matching the pace of the adversarial threat.

In technical implementations, micro-segmentation refers to the division of network environments into isolated zones that limit the lateral movement of compromised processes. In the RAID Framework, this concept is extended to the social information environment: upon detection of a compromise or synthetic operation, automated systems implement a cascade of containment responses scaled to the assessed severity and confidence of the detection. These responses range from enhanced monitoring and shadow-banning of suspect content to full quarantine of compromised accounts and automated notification of potentially affected contacts within targeted networks.

The automation of containment responses is essential because human-speed incident response is fundamentally inadequate against machine-speed propagation. A single viral social media post can reach millions of users within minutes of publication; by the time human analysts have assessed and acted on an alert, the informational damage may be irreversible. The RAID Framework therefore specifies autonomous response capabilities bounded by predefined confidence thresholds and escalation protocols—human oversight is preserved for high-consequence decisions, while routine containment actions are automated.

- **Integration and Governance**

The three pillars of RAID are designed to operate as an integrated system in which each pillar feeds and is fed by the others. Behavioural analytics generates the anomaly signals that calibrate game-theoretic models and trigger micro-segmentation responses. Game-theoretic models generate predictions that direct the attention of behavioural analytics systems and pre-position micro-segmentation capabilities. Micro-segmentation responses generate forensic data that refines behavioural baselines and updates adversarial models.

Governance of the RAID system requires a multi-stakeholder institutional architecture that ensures accountability, prevents abuse, and maintains democratic legitimacy. The framework specifies an independent oversight board with statutory authority, regular public reporting on system operations, mandatory judicial authorisation for monitoring of specific individuals, and robust whistleblower protections for those who identify abuse of the system's capabilities.

Implementation Challenges and Limitations

- **Technical Challenges**

The technical implementation of the RAID Framework faces several significant challenges. The volume and velocity of social media data—hundreds of millions of posts per day across major platforms—creates substantial computational requirements for real-time behavioural analytics. While advances in distributed computing and specialised AI hardware have made this technically feasible, the cost remains prohibitive for resource-constrained national security agencies. The RAID Framework

therefore specifies a tiered implementation model that prioritises monitoring of highest-risk environments and individuals, with comprehensive coverage as a longer-term objective.

The adversarial nature of the threat also creates a fundamental challenge for machine learning-based detection: sophisticated adversaries who understand the detection systems deployed against them can adapt their behaviour to evade detection while remaining operationally effective. This is the "detection arms race" that any automated security system must navigate. The RAID Framework addresses this through diversity of detection approaches—ensuring that no single evasion strategy defeats the entire detection capability—and through continuous retraining of models on current adversarial tactics.

- **Legal and Jurisdictional Challenges**

The cross-jurisdictional nature of social media espionage creates profound legal challenges for defensive operations. Data relevant to national security may be held by platform operators incorporated in foreign jurisdictions, subject to conflicting legal obligations under the laws of their home country and the countries of their users. The legal mechanisms for compelling production of this data—Mutual Legal Assistance Treaties (MLATs), emergency data requests, and emergency legal process—are typically too slow to support real-time defensive operations.

India's evolving legal framework—comprising the Information Technology Act 2000, the Digital Personal Data Protection Act 2023, and the proposed Telecommunications Bill—provides a partial foundation for RAID implementation but requires significant supplementation. Specifically, the framework lacks provisions for real-time threat intelligence sharing between government agencies and platform operators, authorisation for automated defensive responses to detected threats, and explicit legal authority for cross-platform correlation of user data for national security purposes.

- **Institutional and Human Capital Challenges**

The effective implementation of any sophisticated security framework ultimately depends on the human capital available to design, operate, and oversee it. India faces a substantial deficit in both quantity and quality of cybersecurity professionals with the specialised skills required for RAID implementation—AI/ML expertise combined with intelligence tradecraft and legal knowledge. Bridging this deficit requires sustained investment in educational infrastructure, competitive compensation structures that can attract and retain talent against private sector competition, and international partnerships that provide exposure to global best practices.

Policy Recommendations

Based on the analysis presented in this paper and the functional requirements of the RAID Framework, the following policy recommendations are offered for consideration by relevant stakeholders in India's national security ecosystem.

- **Legislative Priorities**
  - **SOCMINT Act:** Enact dedicated Social Media Intelligence (SOCMINT) legislation that provides clear legal authority for government monitoring of social platforms for national security purposes, bounded by robust privacy protections, judicial oversight requirements, and sunset provisions that mandate regular parliamentary review.
  - DPDP Amendment: Amend the Digital Personal Data Protection Act 2023 to create a national security carve-out that enables lawful, judicially authorised cross-platform data correlation for threat detection purposes, with strict data minimisation and purpose limitation requirements.
  - Synthetic Content Law: Develop and enact synthetic content disclosure legislation that requires platform operators to detect and label AI-generated content using technical standards developed in consultation with CERT-In and academic institutions.
- **Institutional Priorities**
  - NSMSC: Establish a National Social Media Security Centre (NSMSC) under the aegis of the National Cyber Security Coordinator, with mandate, resources, and authority to implement the RAID Framework across government networks and to coordinate with platform operators and allied nations' security services.

- SMTISP: Create a Social Media Threat Intelligence Sharing Platform (SMTISP) that enables structured, real-time sharing of threat indicators between CERT-In, NCIIPC, state police cybercrime units, and major platform operators, modelled on the US Cyber Information Sharing and Collaboration Program (CISCP).
- Academic Consortium: Invest in a Cyber Threat Intelligence Academic Consortium comprising IITs, NITs, and central universities to develop indigenous AI-powered detection capabilities and build the human capital pipeline required for long-term RAID implementation.

- **International Cooperation Priorities**
  - Platform Agreements: Pursue bilateral agreements with major platform operators (Meta, Google, Microsoft, X) that establish expedited legal process mechanisms for national security-related data requests, with clear timelines and technical interfaces for automated information exchange.
  - Multilateral Norms: Engage actively in multilateral norm-setting processes—including the UN Group of Governmental Experts (GGE) on Cybersecurity and the Global Partnership on Artificial Intelligence (GPAI)—to advocate for international standards governing state use of AI and deepfake technology for information operations.

Table 2*:* Policy recommendations matrix

Priority	Type	Recommendation	Action Required	Timeline	Lead body
Critical	Legislative	SOCMINT Act — legal authority for social platform monitoring	Parliament to enact dedicated legislation with judicial oversight clause	0–18 months	MeitY · MHA
Critical	Legislative	DPDP Act amendment — national security carve-out	Amend DPDP 2023 for lawful cross-platform data correlation	0–12 months	MeitY · Law Min.
High	Institutional	National Social Media Security Centre (NSMSC)	Establish under National Cyber Security Coordinator with dedicated budget	12–24 months	NSC · NCIIPC
High	Institutional	Social Media Threat Intelligence Sharing Platform	Real-time STIX/TAXII feed between CERT-In, state CERTs, platforms	12–36 months	CERT-In
High	Legislative	Synthetic content disclosure mandate	Require AI-generated content labelling by platform operators	6–18 months	MeitY · TRAI
Medium	International	Bilateral agreements with major platform operators	Expedited legal process for national security data requests	18–36 months	MEA · MeitY
Medium	International	Multilateral norm-setting (UN GGE · GPAI)	Active advocacy for AI/deepfake info-ops standards	Ongoing	MEA
Medium	Institutional	Cyber analytics talent pipeline (IIT/NIT)	500,000 certified professionals target by 2030 via MeitY skill missions	24–60 months	MeitY · Education

Conclusion

The convergence of social media's borderless architecture with the industrialisation of artificial intelligence has produced a threat environment that fundamentally exceeds the adaptive capacity of reactive, perimeter-based security models. The documented operations reviewed in this paper—from GenAI-powered synthetic persona campaigns to deepfake-facilitated financial fraud and systematic LinkedIn targeting of defence personnel—demonstrate that digital espionage has entered a new phase defined by machine speed, machine scale, and machine adaptability.

The national security implications of this transition are not merely technical. They concern the integrity of the information environment in which democratic deliberation occurs, the confidentiality of state and commercial secrets, the reliability of the communications channels through which critical decisions are made, and ultimately the sovereign capacity of nation-states to maintain operational

security in an age when the most sensitive information is separated from adversarial collection only by the strength of social engineering resistance and behavioural vigilance.

The RAID Framework proposed in this paper represents a substantive contribution toward a proactive defensive architecture adequate to these challenges. By integrating AI-powered behavioural analytics, predictive game-theoretic modelling, and automated micro-segmentation into a coherent, continuously adaptive system, RAID offers a conceptual and operational response to the machine-speed threat—one grounded in the recognition that the only effective defence against automation is automation, bounded by human oversight and democratic accountability.

For India, the stakes are particularly high. As one of the world's largest and most rapidly digitising democracies, India presents an attractive and increasingly vulnerable target for social media-enabled espionage and influence operations conducted by state actors with whom it maintains adversarial relationships and by non-state actors whose objectives intersect with those adversaries. The investment required to implement the RAID Framework and the accompanying legislative and institutional reforms is substantial—but it must be measured against the cost of the alternative: continued vulnerability to an adversarial capability that is growing faster than current defensive responses can track.

The transition from a "perimeter-defence" to a "continuous-verification" model of national security is not merely a technical evolution; it is a doctrinal transformation. It requires new legal authorities, new institutional structures, new professional competencies, and new norms of international cooperation. The RAID Framework provides a map for that transformation. The urgency of the threat demands that the journey begin without delay.

References

1. Chesney, R., & Citron, D. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. California Law Review, 107(6), 1753–1820.
2. Hatfield, J. M. (2018). Social engineering in cybersecurity: The evolution of a concept. Computers & Security, 73, 102–113.
3. Macnish, K. (2018). Government surveillance and why defining privacy matters in a post-Snowden world. Journal of Applied Philosophy, 35(2), 417–432.
4. Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. ACM Computing Surveys, 54(1), 1–41.
5. Nimmo, B. (2022). The tactics of disinformation. In N. Wardle & I. Derakhshan (Eds.), Information Disorder: Toward an interdisciplinary framework for research and policymaking. Council of Europe.
6. Omand, D., Bartlett, J., & Miller, C. (2012). Introducing social media intelligence (SOCMINT). Intelligence and National Security, 27(6), 801–823.
7. Prier, J. (2017). Commanding the trend: Social media as information warfare. Strategic Studies Quarterly, 11(4), 50–85.
8. Rid, T. (2020). Active measures: The secret history of disinformation and political warfare. Farrar, Straus and Giroux.
9. Rose, S., Borchert, O., Mitchell, S., & Connelly, S. (2020). Zero trust architecture (NIST Special Publication 800-207). National Institute of Standards and Technology.
10. Rudner, M. (2017). Cyber-threats to critical national infrastructure: An intelligence challenge. International Journal of Intelligence and Counter-Intelligence, 26(3), 453–481.
11. Schneier, B. (2015). Data and Goliath: The hidden battles to collect your data and control your world. W. W. Norton & Company.
12. Taddeo, M. (2017). Cyber security and individual rights, striking the right balance. Philosophy & Technology, 30(4), 461–464.
13. Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. Proceedings of the National Academy of Sciences, 112(4), 1036–1040.

❑❑❑