AI-Enabled Disinformation and National Security: From Research Gaps to Defence Capabilities

Marc-Oliver Pahl (IMT Atlantique), Elie Chedemail (IMT Atlantique), Alexander Loth (Microsoft)

Use Case and Method

Artificial intelligence fundamentally changes the national-security problem posed by disinformation. Large language models (LLMs) can generate convincing misleading narratives at scale [1–3], personalize malicious communication [4,5], sustain interactive influence [6,7], and increasingly operate through autonomous or coordinated agents [8–10]. AI therefore makes economically feasible forms of influence that previously required substantial human resources: individualized interaction, rapid adaptation, large-scale narrative variation, and coordination of automated actors.

For defence and national-security organisations, the problem extends beyond identifying AI-generated content. Influence operations can affect situational awareness, strategic decision making, personnel, strategic communication, force protection, and freedom of action. DISARM accordingly models disinformation as a multi-stage process involving preparation, targeting, content development, dissemination, amplification, and persistence [11].

This paper presents a research-gap analysis and derives corresponding requirements for an AI-assisted national-security defence capability. We analyse two complementary bodies of evidence: a defence-oriented review of LLM-enabled disinformation and an independently constructed corpus of 48 recent studies demonstrating offensive LLM uses and vulnerabilities. We map their coverage against operational needs derived from DISARM: content assessment, target/interaction analysis, coordinated-behaviour detection, temporal/network analysis, and information-source integrity.

Our reasoning follows an explicit chain: **research evidence** $\rightarrow$ **operational coverage** $\rightarrow$ **defence gap** $\rightarrow$ **functional requirement**. The resulting requirements provide the evidence base for a candidate AI architecture and its technology-maturation path. We do not claim an implemented system at this stage.

The central result is a capability mismatch. Research is comparatively mature for AI-generated content, whereas national-security defence increasingly requires capabilities for observing targeted interaction, coordinated AI behaviour, adaptive campaigns, and compromised AI information infrastructures.

The operational use case is **AI-assisted information-environment assessment and decision support**. Individually innocuous messages targeting military personnel, for example, may become significant only when correlated through timing, narrative similarity, account behaviour, personalization, and cross-platform activity.

Research Gap Analysis

Content generation and detection are comparatively mature. LLMs can produce convincing misleading content that may be difficult for humans or automated detectors to distinguish from human-authored information [1–3, 12]. Defence research consequently emphasizes classification, verification, and fact-checking. Yet a “fake/real” decision reveals little about targeting, coordination, propagation, or intervention.

Personalized and interactive influence is insufficiently understood. Studies demonstrate target-specific malicious communication and personalized disinformation [4, 5], while misleading conversational interaction can affect beliefs and false-memory formation [6, 7]. Such influence may target personnel, decision makers, or specific communities while remaining invisible to public-narrative monitoring.

Agentic AI changes the observable from content to behaviour. Research demonstrates autonomous harmful actions, coordinated artificial personas, adversarial social influence, and emergent multi-agent behaviour [8–10, 13, 14]. Adaptive generator–detector loops reinforce the need for temporal, behavioural, relational, and network-level analysis [15, 16].

AI-supported information infrastructure becomes an attack surface. Training-data poisoning and model editing can introduce incorrect knowledge [17–19]; retrieval poisoning can manipulate grounding evidence [20, 21]; and similar attacks affect multi-agent reasoning and fact-checking [22, 23]. Adversaries can therefore manipulate both the information environment and the AI systems used by defence organisations to understand it.

Threat capability	Research maturity	Defence gap	Derived requirement
Generation	Substantial [1, 2]	Artifact decisions lack campaign context	**R1:** Campaign-context assessment
Personalization / interaction	Limited [4, 6]	Targeted influence poorly observable	**R2:** Target/interaction-aware analytics
Agents / coordination	Emerging [8, 10]	Content detectors miss coordinated behaviour	**R3:** Temporal/network coordination analysis
Knowledge integrity	Growing [17, 20]	Trusted AI inputs can themselves be manipulated	**R4:** Provenance/integrity assessment
Composed AI systems	Sparse	Capabilities mostly evaluated independently	**R5:** Multi-source campaign-level fusion

The table is the core output of the analysis: it translates observed research maturity into concrete defence gaps and functional requirements. These requirements, rather than an a priori technology choice, ground the system concept below.

From Research Gaps to Defence Capability

The gap analysis yields five requirements. **R1** requires interpreting artifacts in campaign context; **R2** extends observation to targets and interactions; **R3** adds behavioural, relational, and temporal coordination; **R4** protects the integrity of information feeding AI-assisted analysis; and **R5** fuses these observations because operational systems may combine generation, personalization, retrieval, interaction, and autonomous agents.

These requirements motivate a **human-in-the-loop AI-assisted campaign-level information-environment assessment architecture**. The architecture is a candidate technical realization of R1–R5, not an implemented contribution of the present paper. AI detects, correlates, and prioritizes; analysts validate hypotheses, interpret intent, and decide on operational responses.

- **Detection (R1–R2):** LLM-based information extraction and representation learning identify narratives, entities, targets, interactions, and behavioural signals;
- **Fusion (R3, R5):** *dynamic knowledge graphs* and *temporal Graph Neural Networks (GNNs)* provide candidate mechanisms for correlating actors, content, interactions, and temporal/network relationships;
- **Assessment (R1, R4–R5):** provenance analysis and *neuro-symbolic reasoning* over operational models such as DISARM relate fused evidence to campaign hypotheses, information-integrity risks, confidence, and operational significance;
- **Decision support:** evidence, campaign evolution, and candidate intervention points are presented to human analysts.

The technological development path is therefore explicit: **research gap** $\rightarrow$ **requirements R1–R5** $\rightarrow$ **candidate AI component** $\rightarrow$ **validation objective**. Research and product development are directly coupled: unresolved questions determine which models, data, metrics, and assurance mechanisms must be validated before deployment.

Three challenges condition this maturation. **First**, evaluation must move from artifacts toward campaign evolution, coordination, propagation, and adaptation [1,2,7]. **Second**, operational realism must improve beyond predominantly synthetic, laboratory, red-team, and simulated settings [8–10]. **Third**, AI capabilities must be evaluated in combination rather than independently [4, 20, 24, 25].

Maturity and deployment perspective. The present contribution ends at **requirements analysis and system concept, approximately TRL 2–3**. The maturation path is:

- **TRL 3–4: Experimental demonstrator.** Implement R1–R5 on controlled AI-enabled influence scenarios and establish campaign-level evaluation metrics;
- **TRL 5–6: Relevant-environment validation.** Evaluate multi-source fusion, coordination detection, provenance assessment, and analyst interaction with defence users under representative information flows and adversarial scenarios;
- **TRL 7: Operational prototype.** Integrate validated functions into an information-environment, intelligence, cyber-defence, strategic-communication, or C2-support workflow and evaluate them during an operational exercise or representative mission environment.

The expected operational benefit is **reduced analyst overload, earlier identification of coordinated influence, and evidence-backed prioritization instead of additional raw alerts**. The urgency is immediate: offensive research already demonstrates personalization, autonomous agents, coordination, retrieval poisoning, and attacks against AI-supported information infrastructures [4, 8, 10, 20, 23]. Defence research must therefore close the identified gaps fast enough to translate scientific progress into trustworthy operational capability.

References

[1] I. Vykopal, M. Pikuliak, I. Srba, R. Moro, D. Macko, and M. Bielikova, "Disinformation capabilities of large language models," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024, pp. 14 830–14 847.

[2] A. Williams, L. Burke-Moore, R.-Y. Chan, F. Enock, F. Nanni, T. Sippy, Y.-L. Chung, E. Gabasova, K. Hackenburg, and J. Bright, "Large language models can consistently generate high-quality content for election disinformation operations," *PLOS ONE*, p. e0317421, 2025.

[3] J. Zhou, Y. Zhang, Q. Luo, A. Parker, and M. De Choudhury, "Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–20.

[4] A. Zugecova, D. Macko, I. Srba, R. Moro, J. Kopal, K. Marcincinova, and M. Mesarcik, "Evaluation of llm vulnerabilities to being misused for personalized disinformation generation," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 2025, pp. 780–797.

[5] M. Bethany, A. Galiopoulos, E. Bethany, M. Bahrami Karkevandi, N. Beebe, N. Vishwamitra, and P. Najafirad, "Lateral phishing with large language models: A large organization comparative study," *IEEE Access*, pp. 60 684–60 701, 2025.

[6] P. Pataranutaporn, C. Archiwaranguprok, S. Chan, E. Loftus, and P. Maes, "Slip through the chat: Subtle injection of false information in llm chatbot conversations increases false memory formation," in *Proceedings of the 30th International Conference on Intelligent User Interfaces*, 2025, pp. 1297–1313.

[7] V. Danry, P. Pataranutaporn, M. Groh, and Z. Epstein, "Deceptive explanations by large language models lead people to change their beliefs about misinformation more often than honest explanations," in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2025, pp. 1–31.

[8] A. Tur, N. Meade, X. Lù, A. Zambrano, A. Patel, E. Durmus, S. Gella, K. Stańczak, and S. Reddy, "Safearena: Evaluating the safety of autonomous web agents," in *Proceedings of the 42nd International Conference on Machine Learning*, 2025, no publication DOI found during verification.

[9] M. Gering, Y. Medianyk, A. Sapeha, S. Trendov, K. Karpov, H. Skaskiv, D. Kachan, and E. Siemens, "Botnet simulation and observation framework for ai-driven virtual personas," in *Proceedings of the International Conference on Applied Innovations in IT*, 2025, pp. 75–82.

[10] G. Orlando, J. Ye, V. La Gatta, M. Saeedi, V. Moscato, E. Ferrara, and L. Luceri, "Emergent coordinated behaviors in networked llm agents: Modeling the strategic dynamics of information operations," in *Proceedings of the ACM Web Conference 2026*, 2026, pp. 4805–4816.

[11] S.-J. Terp and P. C. Breuer, "DISARM: A framework for analysis of disinformation campaigns," in *2022 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*. IEEE, 2022, pp. 1–8.

[12] A. Bashardoust, S. Feuerriegel, and Y. Shrestha, "Comparing the willingness to share for human-generated vs. ai-generated fake news," *Proceedings of the ACM on Human-Computer Interaction*, pp. 1–21, 2024.

[13] Z. Lu, G. Lim, and M. Yin, "Understanding the effects of large language model (llm)-driven adversarial social influences in online information spread," in *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 2025, pp. 1–7.

[14] ——, "Large language model (llm)-driven adversarial social influences in online information spread: Risks and interventions," in *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, 2026, pp. 1–19.

[15] C. Tian, Q. Ho, and X. Chen, "A symbolic adversarial learning framework for evolving fake news generation and detection," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025, pp. 12 296–12 310.

[16] C. Santorelli, V. Belmonte, and R. Mastropaolo, "Adversarially enhanced financial misinformation: A comparative analysis of llm- vs. gan-generated content exposing ai moderation vulnerabilities," in *2025 6th International Conference on Artificial Intelligence, Robotics and Control (AIRC)*, 2025.

[17] D. Alber, Z. Yang, A. Alyakin, E. Yang, S. Rai, A. Valliani, J. Zhang, G. Rosenbaum, A. Amend-Thomas, D. Kurland, C. Kremer, A. Eremiev, B. Negash, D. Wiggan, M. Nakatsuka, K. Sangwon, S. Neifert, H. Khan, A. Save, A. Palla, E. Grin, M. Hedman, M. Nasir-Moin, X. Liu, L. Jiang, M. Mankowski, D. Segev, Y. Aphinyanaphongs, H. Riina, J. Golfinos, D. Orringer, D. Kondziolka, and E. Oermann, "Medical large language models are vulnerable to data-poisoning attacks," *Nature Medicine*, pp. 618–626, 2025.

[18] T. Han, S. Nebelung, F. Khader, T. Wang, G. Müller-Franzes, C. Kuhl, S. Försch, J. Kleesiek, C. Haarburger, K. Bressem, J. Kather, and D. Truhn, "Medical large language models are susceptible to targeted misinformation attacks," *npj Digital Medicine*, p. 288, 2024.

[19] C. Chen, B. Huang, Z. Li, Z. Chen, S. Lai, X. Xu, J.-C. Gu, J. Gu, H. Yao, C. Xiao, X. Yan, W. Wang, P. Torr, D. Song, and K. Shu, "Can editing llms inject harm?" in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026, pp. 30 192–30 200.

[20] Y. Anichkov, V. Popov, and S. Bolovtsov, "Retrieval poisoning attacks based on prompt injections into retrieval-augmented generation systems that store generated responses," in *Distributed Computer and Communication Networks*, 2025, pp. 417–429.

[21] R. Sui, X. Yuan, D. Tang, and B. Cui, "Ctrlrag: Black-box document poisoning attacks for retrieval-augmented generation of large language models," in *Proceedings of AAMAS 2026*, 2026, pp. 3597–3599, no publication DOI found during verification.

[22] I. Kraidia, I. Qaddara, N. Al-Shafi, M. Hiari, R. Alzubaidi, and A. Benlecheb, "Rag-induced failures in multi-agent large language model debate," in *2026 1st International Conference on Emerging Technologies and Engineering Systems (ICETES)*, 2026, pp. 326–331.

[23] H. He, Y. Li, B. Zhu, D. Wen, R. Cheng, and F. Lau, "Fact2fiction: Targeted poisoning attack to agentic fact-checking system," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026, pp. 30 943–30 950.

[24] A. Kucharavy, S. Seppey, C. Vallez, D. David, and L. Dolamic, "Low-resource languages llm disinformation is within reach: The case of walliserdeutsch," in *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025, pp. 25 613–25 625.

[25] H.-T. Luong, H. Li, L. Zhang, K. Lee, and E. Chng, "Llamapartialspoof: An llm-driven fake speech dataset simulating disinformation generation," in *2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.