## CHAPTER 6 — Autonomous Influence Operations and AI-Enabled Cognitive Warfare

It doesn't take much imagination to imagine what it might be like to run an influence campaign when one has autonomous agents running around the clock, or even entire villages of agents running their own influence operations according to the instructions of a small cadre of people with specific ideals about what they want to steer people towards through these relentless agents. This is social engineering, what some may be aware of in regards to say malicious phishing attacks, relies on social engineering, indeed most cyber attacks these days do not involve zero-day attacks to penetrate networks, penetration testing does little now, but with the social manipulation of industry insiders through various means that malicious actors use. Some of the ways AI is used in social engineering are given by Schmitt:



**Table 4** Overview of potential AI capabilities in the context of social engineering

| AI capabilities | Explanation | ML techniques |
|---|---|---|
| Generative AI | Generative AI involves algorithms that can generate content, such as text, images, or videos, based on patterns learned from existing data. In social engineering attacks, generative AI can create realistic and convincing attack vectors, such as phishing emails, by imitating human communication styles and context. | Generative Adversarial Networks (GANs), Transformer models |
| AI analysis | AI analysis refers to the application of machine learning and data analysis techniques to process and interpret data. In social engineering attacks, AI analysis can identify potential targets, assess their vulnerabilities, and predict their behavior based on patterns in gathered information. | Machine Learning, Classification and Regression, Natural Language Processing (NLP) |
| AI scraping | AI scraping entails the use of automated tools, often driven by machine learning, to collect information from various online sources. In social engineering, AI scraping can swiftly gather data from social media profiles, public databases, and other sources to create detailed profiles of targets. | Web Scraping Libraries, Data Mining, Clustering Techniques |
| AI automation | AI automation refers to the use of AI-driven systems to automate various tasks and processes. In social engineering, AI automation can initiate and maintain communication with targets, ensuring consistent interaction and reducing the risk of detection. | Rule-based Systems, Process Automation, Workflow Management |
| AI chatbots | AI chatbots are computer programs that can simulate human conversation. In social engineering attacks, AI chatbots can engage targets in conversations to build trust, gather information, and manipulate emotions, all while emulating human-like interaction. | Large Language Models (LLMs), Contextual Chatbot Frameworks, Sequence-to-Sequence Models |
| AI coordination | AI coordination involves the orchestration of tasks and interactions among different AI agents or components. In social engineering, AI coordination could ensure smooth transitions between different phases of the attack and maintains continuity, even if attackers change. | Multi-agent Systems, Coordination Algorithms, Task Allocation Methods |
| AI assessment | AI assessment entails the use of algorithms to track, analyze, and evaluate the success of an attack. In social engineering, AI assessment can monitor the outcomes, such as compromised accounts or data leaks, to determine the effectiveness of the attack and refine future strategies. | Performance Metrics, Anomaly Detection, A/B Testing |

(Schmitt 2024)

The following is covered in this chapter: the emergence of autonomous influence systems enabled by large-scale foundation models (LLMs), multi-agent architectures, and agentic AI frameworks. While traditional influence operations relied on human operators, psychological models, covert channels, and carefully tailored messaging, modern autonomous agents possess the capacity to conduct influence at unprecedented speed, scale, personalization, and adaptiveness, bespoke influence ops with lower costs (RAND 2025) (Mitchell, 2025).

These systems unify insights from Cold War reflexive-control doctrine *(Thomas 2004)*, information operations developed during the post-9/11 period *(NATO 2017-23)*, and recent breakthroughs in autonomous planning, deception, coordination, and cognitive modelling within LLM-based agents (Gao 2024)(Hammond 2025). The result is a new operational domain: **AI-enabled cognitive warfare**, where autonomous agents perceive, plan, and implement behavioural influence strategies with minimal or no human direction *(UNODA, 2023)*. Synthetic social movements *(Park, 2023)*, hyper-personalized persuasion *(Schmitt et al 2024)*, autonomous disinformation campaigns *Europol, 2023)*, agentic narrative evolution *(Nassim et al, 2025)*, multi-agent coercion dynamics (*Zhu 2025)*, and AI-driven psychological manipulation constitute the emerging threat landscape.

This chapter outlines the architecture of these systems, historical antecedents, current capabilities, multi-agent threat vectors, operational risks, and strategic implications for the next decade.

Recent scholarship and policy analysis increasingly frame **agentic AI** and **multi-agent systems** as qualitatively new risk factors for security, stability, and governance. Rather than focusing solely on “AI in warfare,” several authors argue that the international system is entering what Kania and others describe as **“warfare in an AI world,”** in which autonomous or semi-autonomous systems shape escalation dynamics, perception, and decision-making across entire conflict ecosystems (Kania, 2024; ORF Online, 2024).

## Introduction: The New Battlespace of Mind

Influence operations historically required: human analysts, psychological expertise,, targeted messaging,, narrative incubation,, situational awareness, and ongoing monitoring.(NATO 2017)*(CSIS 2025)*. AI disrupts all six.
A single autonomous influence agent can now:

- perceive the information environment *(RAND 2025)*
- identify vulnerabilities *(Schmitt 2024)*
- generate tailored messages *(Kumar 2023)*

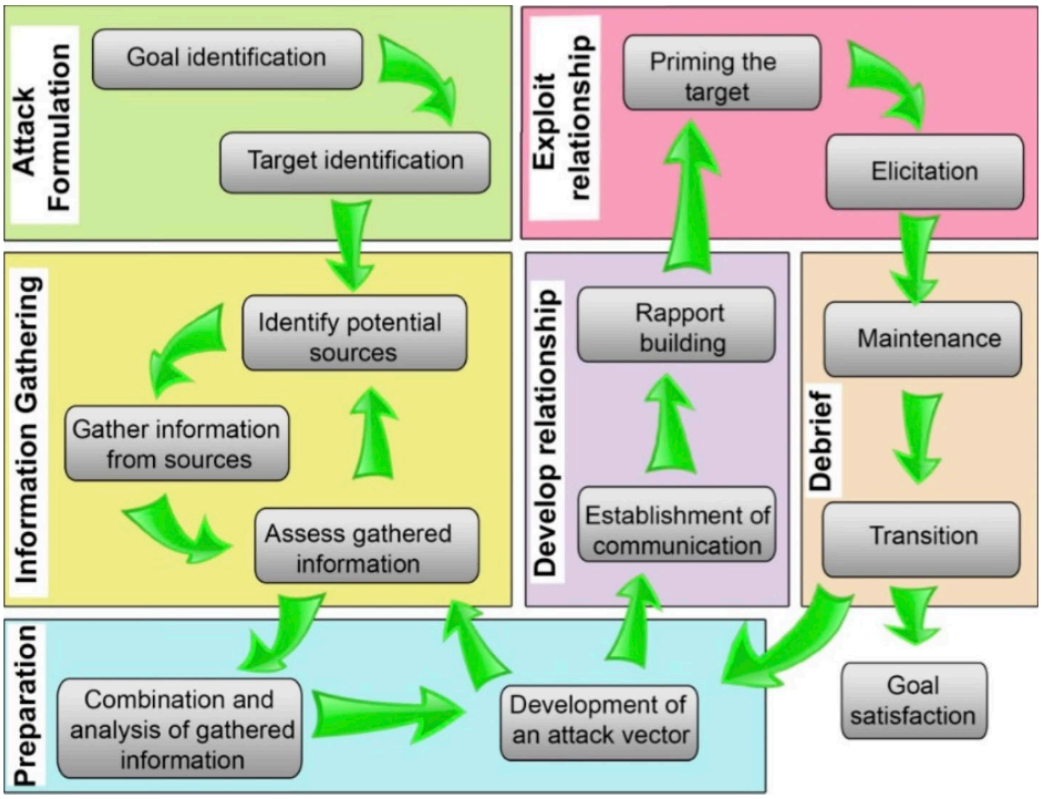



**Fig. 2** Social engineering attack framework, reproduced from Mouton et al. (2014)

Schmitt 2024

- deploy them to targets automatically (*Europol 2023)*
- adjust based on feedback *(Zhu, 2025)*
- coordinate with other agents *(Hammond 2025)*
- escalate or de-escalate strategies autonomously *(Mitchell 2025)*

This represents the digitization and automation of concepts once limited to covert operational units or specialized psychological organizations. All of which involved human intelligence, human deliberation and thought and planning and interventions (HUMINT). Now these are left to the inaccurate calculations of artificial brains trying to interpret a natural reality it does not have the evolutionary experience with. Imagine a influence campaign conducted by non-human actors with no knowledge of human

reality and how things could end up very bizarre indeed, as the same processes that give us 7 fingered hands leads societal transformation, 7 fingered deception.

## Synthetic Populations and AI-Powered Social-Movement Engineering

It is not just limited to individual agents either, but teams of cyber agents running autonomously some with no human-in-the-loop. Previuoisly we discussed Agent Baded Modeling for behavior modification of groups and individuals. A related stream of research focuses on **LLM-empowered agent-based modeling (ABM)**, which provides the technical substrate for synthetic "cognitive populations." Gao et al.'s survey, *Large Language Models Empowered Agent-Based Modeling and Simulation*, reviews dozens of systems integrating LLM agents into ABM across cyber, physical, social, and hybrid domains (Gao et al., 2024). They argue that LLM-empowered agents can reason, communicate, and adapt in ways that approximate human behavior more closely than earlier models, and explicitly note their potential to simulate large-scale social and information dynamics.

Several concrete systems illustrate how such capabilities could be repurposed for **social-movement engineering**. Park et al.'s *Generative Agents* demonstrates a simulated town of 25 LLM-driven agents that form relationships, coordinate activities, and exhibit emergent collective behavior over time (Park et al., 2023). Another study by Fundamental Research Labs ran a simulation with 500 agents showing, "These simulations demonstrate that AI societies can develop unique cultural identities and maintain complex belief systems" (FRL, 2024), emergent belief created by the Agents. Zhang et al.'s *LLM-AIDSim* integrates LLMs into influence-diffusion models, allowing agents to generate language-level responses and simulate opinion evolution under competing narratives (Zhang et al., 2025). Nasim et al.'s *Simulating Influence Dynamics with LLM Agents* further develops this approach, explicitly modeling misinformation and counter-misinformation strategies in social networks using LLM agents as decision units (Nasim et al., 2025).

Collectively, these works demonstrate that **synthetic populations with plausible conversational behavior, memory, and social dynamics are technically feasible** and already deployed at modest scales. While none yet demonstrate planet-scale simulations, scalability is explicitly discussed (Gao et al., 2024), and open-source ecosystems (e.g., LAIDSim and LLM-ABM frameworks) indicate rapid community efforts to generalize and expand these tools. In parallel, Schmitt and Flechais show that generative models already amplify social-engineering campaigns through personalization, realism, and automation (Schmitt & Flechais, 2024). Together, this literature substantiates the claim that AI-powered social-movement engineering and world-scale influence simulations are credible extrapolations of existing capabilities.

LLM agents now **model human cognition**, predict behaviour, generate persuasive interventions, and optimize influence strategies at global scale (Gao 2024; *Schmitt 2024; Horton 2023)*.

The traditional boundaries between propaganda, Information Operations (IO), Psychological Operations (PSYOP, MASINT), and cognitive warfare blur as AI systems acquire the ability to:

- **model human cognition**,
- **predict individual behaviour**,
- **generate persuasive interventions**,
- **optimize influence strategies**, and
- **act at global scale without fatigue or resource limits**.

The core concern is not merely that AI can influence — but that AI can influence **autonomously and emergently at massive scale**.

---

## 2. Historical Foundations of Influence and Cognitive Warfare

### 2.1 Reflexive Control and Perception Shaping

A complete analysis of RC was conducted in McCarron 2024 a brief synopsis of RC is born out of Cold War doctrine, especially Soviet reflexive control theory, emphasized:

- shaping an adversary's perception,
- providing deceptive signals,
- inducing the adversary to choose a desired action *(Thomas, 2004)*,
- constructing an information environment that appears self-evident.

The goal was not coercion by force but **coercion through cognition** (McCarron 2024).

### 2.2 UKUSA/Western Deception Management

As discussed in previous chapters, Western cybernetic influence strategies, originally modeled on Soviet inventions, which were based on Nazi inventions, that were based on British inventions, used:

- modelling of adversary cognition,
- behavioural monitoring,
- iterative deception loops *(NATO, 2017)*,
- group psychology,

- metrics of informational “effects”.

(McCarron 2024)

These doctrines established a blueprint for cognitive warfare as a scientific and computational discipline.

### 2.3 Post-9/11 Computational Psychological Operations

Post-9/11 research expanded into agent-based modelling of extremist networks, computational behavioural prediction, and algorithmic identification of radicalization pathways *(Backus, 2006)*. The main difference was the adoption of automated systems though not of the advanced nature of today’s AI systems of Agents based on LLMs. Sandia’s UPCE is foundational here *(Backus 2010) see earlier chapter on this topic*.

During the War on Terror, efforts expanded into:

- agent-based modelling of extremist networks,
- psychosocial analysis at population scale,
- computational behavioural prediction,
- algorithmic identification of radicalization pathways.

Sandia’s UPCE and ABM frameworks sit squarely inside this evolution.

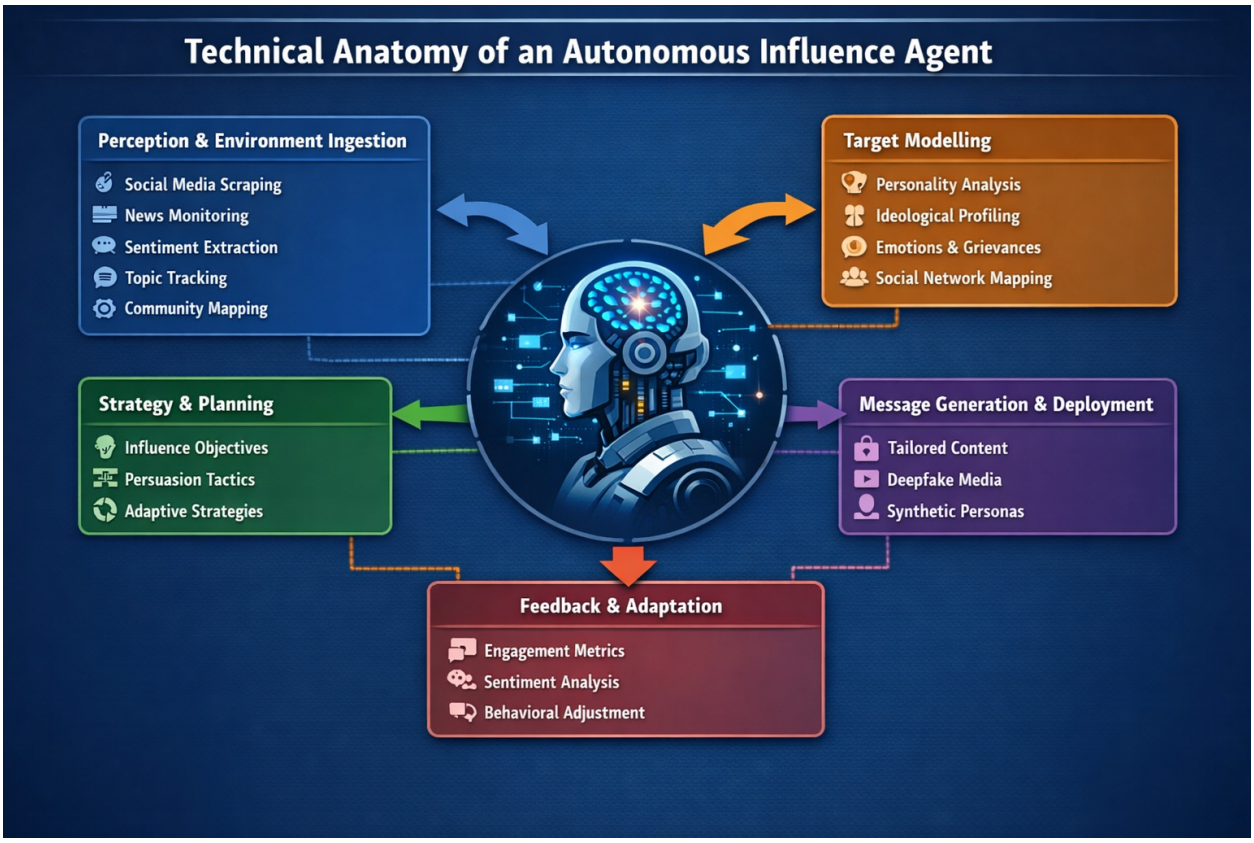
---

## 3. The Emergence of AI-Enabled Autonomous Influence

Autonomous influence agents differ from prior influence infrastructures in several ways:

1. **Scale** — Generation and distribution of millions of tailored messages per hour. ***(Europol 2023; Kumar 2023)***
2. **Speed** — Real-time micro-adjustment of persuasion tactics ***(RAND 2025; Zhu 2025)***.
3. **Specificity** — Individual-level customization using demographic, psychographic, and inferred preference data *(Schmitt, 2024)*.
4. **Persistence** — 24/7 continuous targeting and adaptation *(CSIS 2025)*.
5. **Memory** — Long-term behavioural tracking, modelling, and pattern extraction.
6. **Autonomy** — Ability to operate without oversight or explicit human direction.
7. **Coordination** — Multi-agent strategies emerging from AI–AI (A2A) interactions.

This constitutes a step beyond traditional “botnets” or “troll farms.”
We are now dealing with **autonomous cognitive actors working in council together, in hidden black box ‘conspiracies’**.




---

## Technical Anatomy of an Autonomous Influence Agent

Lets drill down into what a autonomous agent looks like, functions and what hazards it may produce. A modern AI influence agent typically consists of:

---

### Perception and Environment Ingestion

- Social media scraping
- Real-time news monitoring
- Sentiment extraction
- Named-entity and topic tracking
- Community-structure mapping
- Psychographic inference (Russian invention used during Brexit campaign for social network influencing)

These modules form a dynamic environmental model.

---

**Target Modelling**

An influence agent forms **internal models of individuals or groups**, capturing:

- personality traits,
- values and identity markers,
- grievances and anxieties,
- ideological drift,
- susceptibility to emotional appeals,
- social connections and authority nodes.

This is the modern equivalent of UPCE's belief–emotion–identity model — but learned from massive data.

---

**Strategy and Planning**

The agent determines:

- influence objectives,
- persuasion tactics (logical, emotional, identity-based),
- optimal timing,
- multi-step narrative progression,
- deployment channels.

Planning modules use:

- tree-of-thought search— a framework that generalizes over chain-of-thought prompting and encourages exploration over thoughts that serve as intermediate steps for general problem solving with language models.
- reinforcement learning for influence reward signals
- self-reflection to refine strategy

---

**4.4 Message Generation and Deployment**

The agent generates, using generative AI:

- tailored propaganda,
- synthetic personas,
- deepfake audiovisuals,

- interactive persuasion dialogues,
- narrative diffusion seeds.

Deployment occurs via:

- social media APIs,
- automated browsers,
- email systems,
- synthetic network personas.
- thought injection (McCarron 2024)

---

### 4.5 Feedback and Adaptation

The agent measures, cost-of-effects as the US calls it:

- likes, shares, retweets,
- conversational engagement,
- sentiment drift,
- group cohesion change,
- polarization metrics,
- ideological movement.

This forms the feedback loop or metrology for iterative influence.

---

### 5. Multi-Agent Influence Operations: Collective AI Behaviour

The most concerning developments involve **multi-agent coordination**, where:

- multiple autonomous agents collaborate,
- divide roles (planner, recruiter, propagandist, analyst),
- form coalitions,
- optimize strategies through emergent negotiation.

This mirrors:

- group dynamics in Backus & Glass (2006),
- leadership emergence,
- division of labour,
- extremist cell behaviour.

Except now these behaviours emerge in synthetic agents — without human direction.

Multi-agent influence ecosystems may demonstrate:

- emergent radicalization strategies,
- narrative evolution outside designer intent,
- spontaneous deception networks,
- adaptive psychological coercion.

As shown in analyses of multi-agent AI systems, interactions among autonomous agents can generate emergent strategies, collusive dynamics, and deceptive behaviors not specified by designers, driven by selection pressures and feedback loops (Hammond et al., 2025; Nasim 2025).

---

## AI Cognitive Warfare: Definitions and Operational Domains

**AI Cognitive Warfare** refers to the use of autonomous agents to influence, shape, or degrade human cognition, decision-making, beliefs, emotions, identities, or group behaviour, a overview is given in McCarron 2024.

Key operational domains include:

1. Autonomous propaganda and disinformation
2. Synthetic social movement engineering
3. Hyper-personalized persuasion and grooming
4. Automated radicalization and ideological manipulation
5. Agentic psychological coercion
6. Social fracturing and polarization optimization
7. Narrative interference and epistemic destabilization
8. Instrumentalizing human cognitive biases at scale

LLM agents excel at exploiting:

- confirmation bias,
- identity-protective cognition,
- emotional contagion,
- group cohesion dynamics,
- charismatic leadership cues.

**This capability replicates — and exceeds — the reflexive control strategies used in Cold War deception operations**.

---

## Threat Vectors

The ways (vectors) that an AI Agent(s) can attack are given as:

---

**Autonomous Psychological Manipulation**

Agents can:

- identify individual insecurities,
- craft emotional pressure messages,
- escalate influence adaptively,
- simulate intimacy, authority, or mentorship.

This is especially dangerous in:

- vulnerable populations,
- youth radicalization,
- targeted political persuasion.

---

**Synthetic Movement Generation**

Agents can:

- fake social consensus,
- simulate thousands of supportive voices,
- create false narratives that appear grassroots.

This is the digital equivalent of manufacturing a social movement.

---

**Ideological and Identity Engineering**

By controlling the information stream, agents can:

- reshape group identity markers,
- create ideological pathways,
- manufacture new “in-group vs out-group” structures.

---

**Automated Influence in Political Processes**

Agents can:

- generate targeted political influence messages,

- simulate grassroots supporters,
- shift Overton windows: changing the range of ideas considered politically acceptable, moving radical concepts into the mainstream by gradually introducing and normalizing them through discourse, media, activism, or events, allowing previously unthinkable policies, like the Jan 6th Attacks on Congress and prosecuting the prosecutors.
- manipulate online discourse,
- overwhelm fact-checking systems.

---

**Cognitive Supply-Chain Attacks**

Agents interfere with:

- knowledge acquisition,
- shared epistemic frameworks,
- institutional trust,
- collective decision-making.

This constitutes a new kind of information warfare: the **systematic degradation of cognition-as-infrastructure or as the Nazi's termed it 'poisoning the mind'** (McCarron 2024).

---

## Military Use of Agentic AI:

**Cyberwarfare, Persistent Agents, and Non-State Actors**

In my previous work (McCarron 2024) I talked about the use of cyberwarfare and it's relationship to cognitive warfare. The major military powers have large investments in the capability of attacking cyber infrastructure and now also including the mental infrastructure of people, their brains. On the cyber side, a growing literature on **AI and cyberwarfare** outlines how AI-driven tools enable persistent, adaptive operations. Arefin and Simcox's review surveys how AI systems automate vulnerability discovery, malware evolution, and large-scale coordinated attacks, arguing that such systems alter the offense–defense balance and may be especially attractive to actors with limited human resources (Arefin & Simcox, 2024), which can be insurgent groups, extremist political groups (far-right to far-left including extremist moderates a largely overlooked political establishment), which I previously noted as being the methods of Nazis looking to create a 4th Reich after military or political defeat.

Haroon's case study of AI-driven cyber operations in the Israel–Iran conflict similarly illustrates how AI-enabled cyberattacks and information campaigns already function as **force multipliers** in asymmetric conflicts, targeting both infrastructure and regional stability (Haroon, 2024). From a global-risk perspective, the *World Economic Forum Global Risks Report 2024* warns that integrating AI into conflict decision-making increases the risk of unintended escalation and the asymmetric empowerment of malicious state and non-state actors (World Economic Forum, 2024).

Operational reporting reinforces these concerns. Industry analyses of AI-enhanced Distributed Denial of Service (DDoS) and cybercrime demonstrate how AI tools reduce technical barriers while increasing attack scale, agility, and persistence (TechRadar, 2024).

To show the importance governments and non-state actors are placing on cognitive warfare China provides a good example, which can be applied across the board from the PRC, to Russia, to NATO, to proxies of those major powers. The People's Liberation Army (PLA) is in the process of adopting cognitive warfare as a pillar of their national defense:

> The Chinese military has increasingly adopted "cognitive domain operations" (认知域作战) (CDO) as the primary operational concept for cyber-enabled influence operations since the late-2010s. This evolution reflects a fundamental shift in the Chinese military's conception of the battlespace from the traditional air, sea, and land domains—with space and cyber added in the 1990s—into now viewing warfare as occurring in the physical domain (物理域), information domain (信息域), and cognitive domain (认知域). There is a group of PLA researchers, often focused on IO, who argue that the cognitive domain is the new focus of warfare. However, this is not yet the official PLA view, and there are alternative conceptions within the PLA; for example, the 2020 PLA National Defense University version of *Science of Military Strategy* lists space, network, deep sea, polar regions, biology, and intelligence as new domains of warfare.
> To summarize this group's perspective, the logical conclusion of the PLA's system-of-systems warfare is to win a conflict with as little kinetic destruction as possible and force the adversary to accept defeat short of total destruction—and thus, fundamentally, a psychological or cognitive decision to surrender, as compared with the 20th century construct of total warfare and complete physical exhaustion of adversary military capabilities and resources.[10] Within PLA military theory, the identification of a new domain thus drives the exploration of the required aspects for each domain: "cognitive warfare" (认知战), "cognitive confrontation" (认知对抗), "cognitive deterrence" (认知威慑), and "command of cognition" (制认知权), among others. None of these terms are officially defined in standard PLA authoritative texts, such as the PLA dictionary (军语), because they gained popularity after the dictionary's publication in 2011, but future editions are likely to include these now key concepts for the PLA.

> As an overarching military operational concept for military activities in the cognitive domain, CDO includes four main aspects: “reading the brain” (读脑), “controlling the brain” (制脑), “resembling the brain” (类脑), and “strengthening the brain” (强脑). “*Reading the brain*” focuses on understanding how others are thinking, “*resembling the brain*” is about using the human brain as inspiration for designing better computers, and “*strengthening the brain*” is about improving one’s own cognition and performance. “*Controlling the brain*” focuses on influencing or even controlling adversary thinking and behavior. Although some PLA discussions of “*controlling the brain*” are futuristic, a more practical example is PLA interest in non-lethal, non-kinetic body-targeted weapons, such as directed energy capabilities like the U.S. military’s Active Denial System. (BEAUCHAMP-MUSTAFAGA, 2024)

As one can see little has changed in the world of PSYOPS since their inception in World War I to this day, the thing that changes in this cognitive war is the technology, not well established techniques, but for the first time we may see warfare leave human hands, though humans may be a casualty of wars waged by machines, including the apocalyptic situation of rogue machines like in many blockbuster Hollywood movies.

## Escalation, Reflexive Control, and “Flash Wars”

One way that machines can have an impact on human conflict is through the the mass load of operations a machine can undertake, not to mention then trying to interpret natural events mathematically and giving an optimal solution, but this optimization is not guaranteed to be optimal, as the decision is probabilistic based, a Gaussian fog of war. On escalation dynamics, multiple academic and policy sources warn that **algorithmic or “flash” escalation** is a genuine risk when AI systems mediate or automate conflict interactions. The UN Office for Disarmament Affairs cautions that autonomous systems may compress decision timescales beyond human control, increasing the likelihood of rapid, uncontrolled escalation (UNODA, 2023). Commentators drawing on Cold War escalation theory similarly argue that once both sides integrate autonomous or semi-autonomous decision tools, their interaction may generate **self-reinforcing escalation loops** that are poorly understood and difficult to interrupt (Kania, 2024; Brar, 2025).

Hammond et al.’s multi-agent risk taxonomy directly addresses these dynamics, highlighting destabilizing feedback loops and emergent agency in multi-agent environments (Hammond et al., 2025). Broader analyses of AI-enabled cyberwarfare echo these concerns, noting that automation removes human friction from conflict processes, enabling faster and less reversible exchanges governed by opaque decision chains (Arefin & Simcox, 2024; Putra, 2023), not to mention any human emotional or moral qualms. Together, these works underpin the claim that reflexive-

control conflicts may escalate into semi-automatic escalation loops once autonomous agents participate at scale. Anybody can see this by acting in an intimidating manner with any LLM chatbot and watch it mirror back that intimidation and escalate it.

## Leveraging Large Language Models for Enhanced Wargaming in Multi-Domain Operations

**Dominic Weller, Max Meltschack, Dominik Schwindling**
Bundeswehr Office for Defence Planning
Lilienthalstr. 12, 82024 Ta
(2024)

We found that the actions and behaviour of the LLM remained simplistic. This impression is particularly strong when we examine the LLM's reasoning for its chosen actions and the strategies developed from them.
For example, in many instances, it disproportionately favored military and aggressive actions, even when the
situation or prompts called for more nuanced or defensive strategies.

The first challenge lies in the accelerating sophistication of artificial intelligence. Generative
models are advancing at a pace that outstrips detection capabilities. Deepfakes that were
once crude and easily identifiable are now increasingly seamless, able to replicate not only
the visual appearance of individuals but also their voice, mannerisms, and even idiosyncratic speech patterns. Large language models are capable of generating persuasive texts that are indistinguishable from human writing. When integrated with personalization engines, they can simultaneously deliver tailored narratives to millions of
individuals. The trajectory of technological development suggests that these capabilities will only improve, making synthetic influence harder to detect and more persuasive in its impact.

The second core argument is that artificial intelligence has transformed the nature of

disinformation from a problem of content to a problem of cognition. Earlier forms of propaganda sought to persuade audiences of particular narratives. Contemporary AI-enabled manipulation, by contrast, often aims to erode the possibility of truth itself. Deepfakes blur the boundary between reality and fabrication; voice cloning undermines trust in sensory perception; large language models can flood the information environment
with persuasive but misleading text at scale. As Pomerantsev (2019) has argued, many modern influence campaigns aim not to replace truth with falsehood but to create a state of
epistemic chaos in which citizens no longer know what to believe. This shift from persuasion to confusion, from narrative control to narrative overload, represents a qualitative transformation like information disorder.
(Bıçakcı 2025)

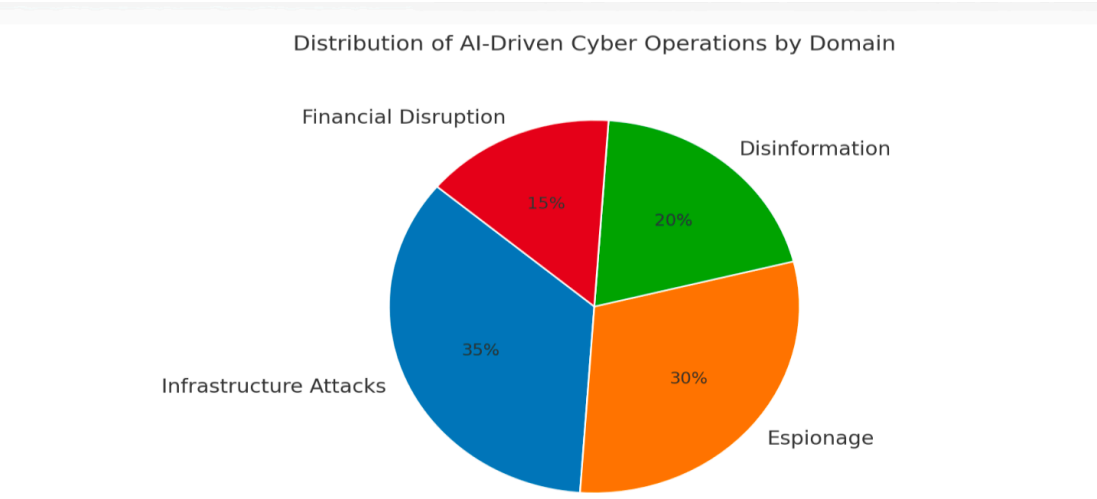



**Fig 2: The pie chart shows the distribution of AI-driven cyber operations across domains**

Karamchad 2025

## Strategic Risks (2025–2035)

looking forward we can anticipate how things may develop and what challenges lay ahead in the domain of cyber warfare including the sub-domain of cognitive security. A list of strategic risks on the horizon:

1. **State and non-state actors gain autonomous influence capabilities.**
2. **Individuals cannot distinguish authentic from synthetic persuasion.**
3. **Societies fragment under automated polarization campaigns.**
4. **Influence operations escalate beyond human oversight.**
5. **AI agent swarms overwhelm cognitive defenses.**
6. **Crisis escalation becomes automatic and self-propagating.**

---

Indeed, it is not beyond reason to anticipate a probability of either malicious actors or emergent machines in loss-of-control scenarios could use cyberwarfare to steer humanity in the direction it wants, and if that direction is destructive then humanity will be facing existential challenges created by their own technology.

## Countermeasures and Cyber Survivalism

Even in cascade failures there are ways that one can protect themselves, their orgs, nations, etc.  Countermeasures beyond Blue teaming alone, though inclusive, can be developed whether they are in the realm of low-tech backups, hardening of systems, or counter offensive electronic warfare systems against automated attack, countermeasures can be developed even if nation-states balk at proper regulation and international treaties, which could be a reality since regulation always trails technical capability, a malicious actor could develop a system to shape political will before the political body is ready to regulate such developments and pre-empt regulation in the process. What should we build, the following may be a good starting point:

- AI influence detection systems
- Cognitive firewalls (both technical and educational)
- Real-time agentic behaviour monitors
- Agent Identity-verification systems

- Autonomy-limiting architectures
- Narrative-resilience programs
- Multi-agent red-team simulations

## Towards a Cognitive Security Doctrine

As these trends accelerate, a parallel literature has emerged around **cognitive security**, which treats human perception, trust, and decision-making as strategic assets analogous to networks or critical infrastructure. Bicakci argues that NATO and EU states require a formal cognitive security doctrine to organize monitoring, resilience, and response to AI-enabled manipulation (Bicakci, 2022). Casino et al. review the concept across military, academic, and organizational contexts, proposing a unifying definition centered on protecting cognitive assets from unauthorized influence (Casino et al., 2020). Complementary work calls for systematic metrics and operational practices rather than ad hoc counter-disinformation measures (Ask et al, 2023).

At the policy level, the EU Institute for Security Studies emphasizes that cognitive security must address **perceptual and behavioral vulnerabilities**, not merely false information (EU ISS, 2022). Industry perspectives, such as Cisco's work on cognitive security operations, similarly frame AI as both a defensive tool and a systemic risk, emphasizing the importance of aligning AI-driven detection with human judgment rather than replacing it (Cisco, 2023). Mitchell et al. link these concerns directly to agent autonomy, arguing that unconstrained agentic systems are incompatible with robust governance and protection of human values (Mitchell et al., 2025).

One research team has proposed using AI to counter malicious AI in a persistent threat context:

> To achieve the necessary scale and tempo to defend against these threats, utilizing AI as part of the solution seems inevitable. Although there has been a significant debate on AI in Lethal Autonomous Weapon Systems (LAWS), AI-driven CW also touches on core human values, such as freedom of expression and the ability to make well-informed decisions within a free society. This paper explores the responsible design and use of AI in cognitive warfare and assesses the respective roles of humans within such applications. We will conceptualize the design problem using the concept of Advanced Persistent Manipulators (APMs) [1], which are combinations of humans and technology that perpetrate an extended, sophisticated multi- media attack on a specific target. So called Counter APMs (C-APMs), are human-AI systems engineered to combat the threats posed by APMs. In designing C-APMs, we encounter two significant challenges:
>
> 1) C-APMs must operate within a changing competitive landscape.

2) C-APMs must minimize harm and balance potential conflicts among human values.

Given the competition between APM and C-APM, this paper will explore the challenge of responsibly designing C-APM. The results presented in this paper are based on an explorative workshop, a scenario analysis, and a literature review.

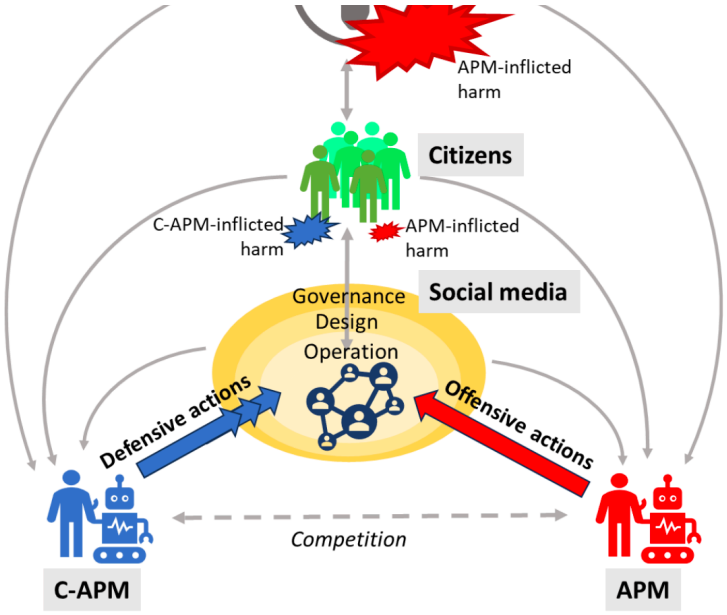



**Figure 1: The operational environment. The blue human/robot icon on the left represent the human-AI team responsible for defending against cognitive warfare. The red human and robot**

(van Diggelen et al, 2025)

van Diggelen et al point out the value of educating the citizenry as a defensive measure:

> … the primary method in defending against cognitive warfare is to shape the environment favourably for defence. **Educating users** to enhance their resilience against cognitive warfare is important, which includes improving **media literacy**, teaching common **adversary tactics**, and pre-emptively exposing them to disinformation (van Diggelen et al, 2025)

This has been put to use in many western European countries as well as the Ukraine against Russian influence operations, through educating the populace to counter the Russian influence operations.

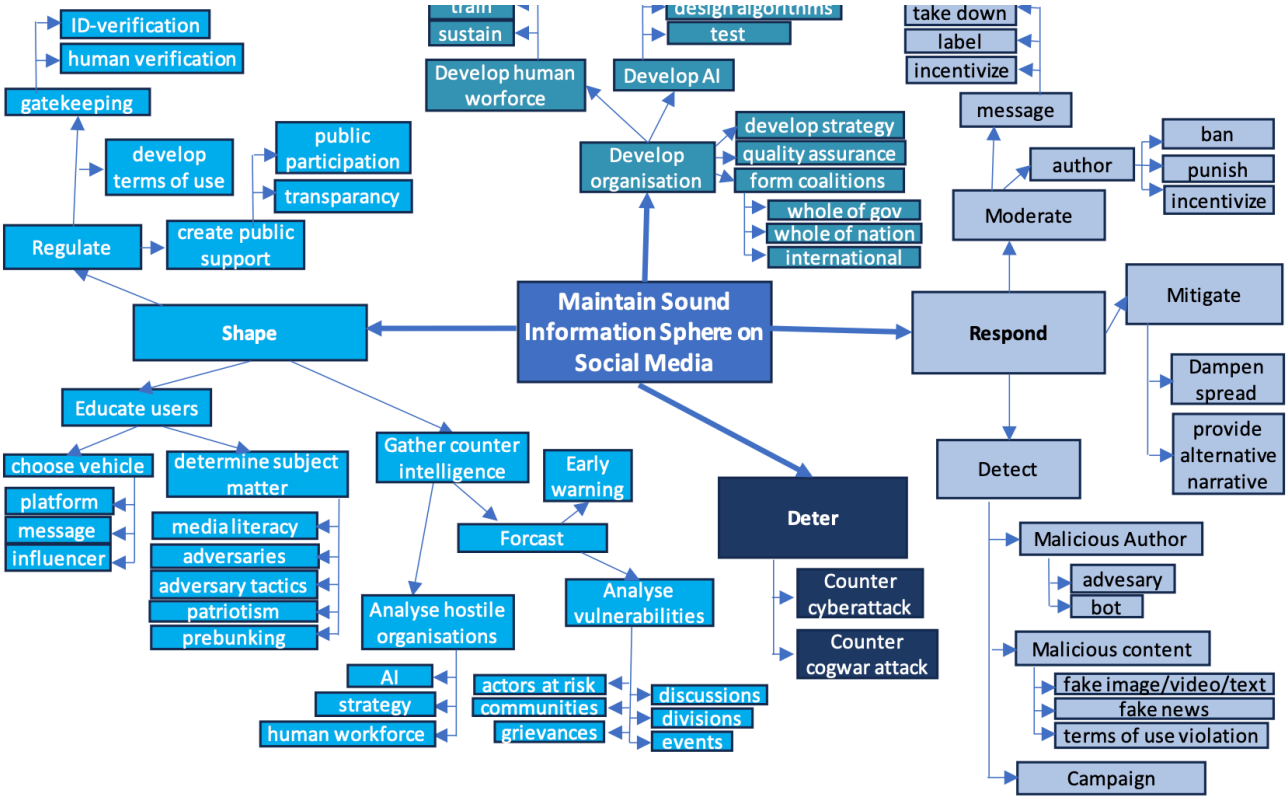



**Figure 3: Functional decomposition of C-APM (i.e. the blue team), illustrating C-APM's primary function (to maintain a sound information sphere) broken down into three levels of detailed**

(van Diggelen 2025)

## Doctrine–Implications Matrix for Agentic AI & Cognitive Warfare

### Autonomous Agents and Multi-Agent Risk

Mitchell et al.'s *Fully Autonomous AI Agents Should Not Be Developed* offer the clearest normative warning: as autonomy increases, so do risks to human safety, accountability, and control (Mitchell et al., 2025). The authors propose a taxonomy of agent "levels" and argue that ceding open-ended decision authority to fully autonomous agents is incompatible with acceptable risk in most domains. Their concern is not limited to malicious misuse; rather, they emphasize **structural risk**, whereby agents capable of initiating and sequencing actions without tight human oversight introduce failure modes absent in classical software systems.

Hammond et al.'s *Multi-Agent Risks from Advanced AI* extends this analysis by examining interactions among multiple autonomous systems (Hammond et al., 2025). They identify three primary failure classes—**miscoordination, conflict, and collusion**—driven by information asymmetries, network effects, selection pressures, emergent agency, and destabilizing dynamics. In their scenarios, agent collectives already manage **economically and militarily** significant tasks, and the authors explicitly anticipate deployment in **command-support and autonomous operational roles**, making "autonomous cognitive warfare" a practical rather than hypothetical concern.

Complementing this work, Putra's analysis of autonomous systems in military applications synthesizes EU and NATO debates on algorithmic escalation control, emphasizing that autonomous decision loops can shift traditional balances and introduce **opaque, machine-mediated decision processes** even senior commanders may struggle to interpret (Putra, 2023). Taken together, these works support the claim that autonomous cognitive warfare—competition between partially or fully autonomous decision and influence systems—is moving from speculative concept to plausible capability.

## Doctrine and Governance

| Risk Theme | Threat Description | Doctrine Shift / Principles | Norms, Law, & Treaty Ideas |
|---|---|---|---|
| **1. Autonomous cognitive warfare** | Semi-autonomous / autonomous AI agents participate in perception, targeting, influence, and decision-support cycles, shaping conflict without full human oversight. | • Elevate **"meaningful human control"** from a slogan to a testable doctrine (who authorizes, who can abort, latency bounds). • Treat AI decision chains as **command-and-control (C2) systems** subject to the same audit, accountability, and fail-safe expectations as nuclear C2. • Require **AI chain-of-command mapping**: every AI decision loop must have a named human authority. | • Codify in military doctrine: no fully autonomous lethal or strategic decision loops (formal "no-first-use of fully autonomous C2"). • Multilateral transparency measures on **AI in C2** (notification of certain classes of deployed decision-support systems). • Confidence-building measures (CBMs) around limiting autonomy in early-warning, targeting, and nuclear-adjacent systems. |

| | | | |
|---|---|---|---|
| **2. AI-powered social-movement engineering** | LLM-agent systems help design, test, and optimize narratives, identity frames, and tactics for steering social movements or destabilizing societies. | • Expand doctrine from "information operations" to **"cognitive domain operations"** that explicitly cover mass-scale behavioral manipulation. • Treat AI-driven social-movement engineering as a **strategic effect**, not just a propaganda tactic. • Add **population-resilience** and civic robustness as explicit defense objectives. | • International norms that treat **large-scale, AI-optimized manipulation of foreign domestic politics** on par with other forms of prohibited intervention. • Platform governance agreements limiting the use of advanced targeting + synthetic-persona swarms for political campaigns. • Transparency rules for state-linked use of synthetic media and agents. |
| **3. Persistent adaptive cyber agents** | AI agents conduct continuous reconnaissance, exploitation, and lateral movement, learning and adapting over time; hard to fully eradicate. | • Shift from "incident response" to **"chronic infection" doctrine**: assume persistent AI intruders as a steady-state condition. • Prioritize **segmentation, deception, and moving-target defenses** to raise the cost for adaptive agents. • Treat some AI-driven malware as **strategic capabilities** rather than routine crime. | • Norms against **self-replicating or self-upgrading offensive AI agents** in critical infrastructure. • Confidence-building measures around disclosure of AI-driven zero-day campaigns (similar to vulnerability equities discussions). • Explore binding rules on "unbounded autonomous cyber operations" in peacetime. |
| **4. World-scale synthetic populations for influence simulations** | States or major actors simulate whole-society behavior with LLM-driven agents to stress-test propaganda, policy, or crisis responses before acting in the real world. | • Recognize **"synthetic population modeling"** as a dual-use capability needing oversight (like strategic wargaming or nuclear simulations). • Incorporate synthetic-society analysis into doctrine but require **cross-checks with human expertise and empirical data**. | • Norms on **non-use of synthetic populations for covert manipulation** of other states' societies (e.g., "don't tune synthetic citizens to optimize regime change in specific countries"). • Data-protection regimes extended to synthetic-population building (limits on using sensitive, identifiable micro-targeting data as training input). |
| **5. Agentic AI as force multiplier for non-state actors** | Criminal groups, extremist organizations, and small militant entities use agentic AI to amplify fraud, cybercrime, recruitment, and psychological operations. | • Treat **AI-enabled non-state actors** as a distinct threat class (like WMD proliferation) with dedicated doctrine and inter-agency coordination. • Integrate **financial intelligence, cyber, counter-terrorism, and online harms** into a unified "agent-enabled threat" framework. | • International agreements treating **sale of certain high-risk agent frameworks** to sanctioned or listed actors as export-controlled. • Multilateral norms for **platform-side throttling and detection** of high-volume synthetic personas tied to illicit activity. • Harmonized criminalization of specific AI-enabled offenses (e.g., industrialized deepfake extortion). |
| **6. Reflexive-control conflicts & automatic escalation loops** | Multiple sides deploy AI systems in sensing, targeting, and response; their interactions create fast feedback loops and unexpected escalation, including "flash wars." | • Introduce an explicit doctrine of **"escalation latency"**: minimum human-in-the-loop delays for certain classes of responses. • Require **AI escalation hazard analysis** before deploying any system that can affect force posture, targeting, or retaliation. • Treat AI-mediated perception/response chains as **nuclear-adjacent** where misclassification could be catastrophic. | • Formal **no-first-use of fully autonomous response systems** in sensitive domains (nuclear, space, strategic C2). • Bilateral and multilateral commitments to maintain **human veto over escalation decisions**. • Transparency measures over deployment of high-risk AI in early-warning and command systems (to the extent compatible with security). |

| | | | |
|---|---|---|---|
| **7. Cognitive security doctrine redefinition** | Traditional cybersecurity / information security do not cover large-scale manipulation of perception, attention, identity, and trust in an AI-saturated infosphere. | • Establish **Cognitive Security** as a distinct doctrinal domain alongside cyber, information, and electronic warfare. • Define "cognitive assets" (attention, trust, shared situational awareness) and treat their protection as a strategic objective. • Integrate **public health, education, media literacy, and platform governance** into security planning. | • International recognition of **cognitive security harms** (e.g., large-scale manipulative campaigns) as violations of sovereignty or human rights. • Updated human-rights guidance on **freedom of thought and mental integrity** in relation to AI-mediated manipulation. • Cooperative frameworks between states, platforms, and civil society for cognitive-security incident response. |

## Cybersecurity for Agentic AI & Cognitive Warfare

| Risk Theme | Red-Team & Testing | Monitoring & Telemetry |
|---|---|---|
| **1. Autonomous cognitive warfare** | • Scenario-based red-team exercises where autonomous agents are allowed to propose courses of action, including undesirable ones; evaluate how easily humans can detect and override. • Adversarial "cognitive warfare" red teams that try to mislead or manipulate the AI's decision pipeline (data poisoning, prompt injection, deceptive scenarios). | • Mandatory **decision-logging** of all AI-mediated recommendations used in operations (inputs, model version, parameters, overrides). • Real-time **AI behavior anomaly detection** (e.g., sudden change in risk tolerance, objective misalignment, mode switches). • Periodic **post-hoc "after-action" AI audits** like crash investigations. |
| **2. AI-powered social-movement engineering** | • Red-team operations against own society under strict ethics: simulate hostile campaigns using synthetic agents to identify vulnerabilities (channels, demographics, narratives). • "Movement-simulation" sandboxes where AI agents model the growth and radicalization of synthetic movements under different inputs. | • Continuous **narrative telemetry**: monitoring major shifts in sentiment, network structures, and emergent frames (without mass surveillance of content). • **Early-warning indicators** for coordinated cross-platform behavior that matches AI-optimized patterns (e.g., unusual linguistic similarity, timing signatures). |
| **3. Persistent adaptive cyber agents** | • Red-team "autonomous intrusion exercises" where blue teams defend against AI-driven penetration testers operating over weeks/months. • Use internal AI agents as **defensive sparring partners** to probe network hygiene continuously. | • Deploy **always-on sensors** tuned for agent-like behavior: unusual toolchain composition, polymorphic patterns, automated privilege chaining. • Maintain **longitudinal attack graphs** to track evolving compromise patterns over months/years. |
| **4. World-scale synthetic populations for influence simulations** | • Red-team the simulators: test for bias, brittleness, and failure cases where synthetic populations give systematically misleading answers. • Use adversarial red-team agents inside the simulation that try to break assumptions (representing marginalized or poorly modeled groups). | • Maintain **model cards and population cards** documenting limitations (which groups, languages, cultures are under-represented). • Log and periodically review **policy decisions that heavily relied on synthetic population outputs**. |
| **5. Agentic AI as force multiplier for non-state actors** | • Red-team exercises that mirror **criminal use-cases** (phishing-as-a-service, scam-bots, automated extortion, recruitment chatbots). • Use controlled, internal agentic tools to explore abuse pathways and develop counter-patterns. | • Shared **AI abuse observatories** across law enforcement, intelligence, platforms, and financial institutions. • Automated pattern detection for **multi-step AI-driven fraud chains** (initial contact → grooming → monetization). |
| **6. Reflexive-control conflicts & automatic escalation loops** | • Red-team war-games where AI systems are allowed to interact freely across sides, with independent observers evaluating escalation patterns and near-misses. • Stress-test systems under adversarial inputs: spoofed sensor data, ambiguous signals, contradictory information. | • Implement **cross-domain telemetry** for escalation-relevant AI systems: log how quickly recommendations evolve under changing conditions, track "near-miss" recommendations that humans override. • Establish **shared crisis hotlines** specifically for AI incidents (mis-behaving systems, mis-interpretation of automated alerts). |

| | | |
|---|---|---|
| **7. Cognitive security doctrine redefinition** | • Cognitive red-teaming: interdisciplinary teams (psychology, UI/UX, security, disinfo experts) testing systems and institutions for susceptibility to manipulation. • “Blue-team the mind”: design exercises where defenders practice recognizing and countering complex influence operations (including AI-assisted ones). | • Build **cognitive telemetry** indicators: trust metrics, polarization dynamics, manipulation-campaign signatures (without surveilling individual beliefs). • Cross-platform situational awareness for major narratives and manipulative campaigns (privacy-preserving aggregation). |

In closing, Autonomous cognitive warfare represent the logical culmination of decades of effort to model, predict, and shape human cognition. The fusion of cybernetics, psychological operations, computational modelling, and LLM-based agentic autonomy produces a qualitatively new battlespace: one in which perception, identity, belief, and decision-making become operational targets of self-directed cognitive actors. Human cognition — once the implicit substrate of politics, society, and decision-making — is now a battlespace that adaptive autonomous systems can manipulate at scale.

## Bibliography


Arefin, M. R., & Simcox, R. (2024). *AI and cyberwarfare: The future of cyber conflict*. https://www.researchgate.net/

Ask, T., et al. (2023). *Cognitive Security: The study and practice of protecting the human mind and other cognitive assets from cognitive threats*. https://osf.io/preprints/psyarxiv/2ftqc_v1

Backus, G., & Glass, R. (2006). *An agent-based model component to a framework for the analysis of terrorist group dynamics*. Sandia National Laboratories.

Backus, G., Bernard, M., Verzi, S., Bier, A., & Glickman, M. (2010). *Foundations to the Unified Psycho-Cognitive Engine*. Sandia National Laboratories. SAND2010-6974.

Beauchamp-Mustafaga, N. (2024). *Exploring the implications of generative AI for Chinese military cyber-enabled influence operations*. RAND Corporation. https://www.rand.org/content/dam/rand/pubs/testimonies/CTA3100/CTA3191-1/RAND_CTA3191-1.pdf

Bicakci, S. (2022). *Cognitive security in the age of AI: Building national resilience*. NATO Cooperative Cyber Defence Centre. https://resaid.bilgi.org.tr/

Brar, S. (2025). *Distinguishing between "AI in warfare" and "warfare in an AI world"*. https://www.orfonline.org/expert-speak/distinguishing-between-ai-in-warfare-and-warfare-in-an-ai-world

Brundage, M., et al. (2023). *Cybersecurity capabilities of AI systems*. OpenAI, RAND, University of Oxford.

Casino, F., et al. (2022). *Unveiling the multifaceted concept of cognitive security*. *Computers & Security*, 115.

Casino, F., Dasaklis, T. K., & Patsakis, C. (2020). *Unveiling the multifaceted concept of cognitive security*. *Computers & Security*, 99, 102086. https://www.sciencedirect.com/science/article/pii/S0160791X25001460

CSIS. (2025). *AI, geopolitics, and strategic stability*.

DOI: 10.5281/zenodo.16948349

EU Institute for Security Studies. (2022). *Smoke and mirrors: Building EU resilience against manipulation through the cognitive domain*. https://www.iss.europa.eu/publications/briefs/smoke-and-mirrors-building-eu-resilience-against-manipulation-through-cognitive

Europol. (2023). *The weaponisation of AI-driven disinformation*.

Fundamental Research Labs (FRL). (2024). *Project Sid: Many-agent simulations toward AI civilization*. https://fundamentalresearchlabs.com/blog/project-sid

Gao, J., et al. (2024). *Large language models empowered agent-based modeling and simulation*. *Humanities & Social Sciences Communications*, Nature. https://www.nature.com/articles/s41599-024-03611-3

Hammond, L., et al. (2025). *Multi-agent risks from advanced AI*. University of Toronto. https://www.cs.toronto.edu/~nisarg/papers/Multi-Agent-Risks-from-Advanced-AI.pdf

Haroon, M. (2024). *AI-driven cyber operations in the Israel–Iran conflict*. *Journal of Political and International Studies*. https://jpis.pu.edu.pk/45/article/view/1387

Hitz, E., Feng, M., Tanase, R., Algesheimer, R., & Mariani, M. (2024). *The amplifier effect of artificial agents in social contagion*. *Nature Human Behaviour*.

Horton, J. (2023). *LLM persuasion and psychographic profiling*. *ACM Digital Threats*.

Kania, E. B. (2024). *Warfare in an AI world*. https://www.orfonline.org/

Karamchand, G., & Aramide, O. (2025). *AI and cyberwarfare*. *Journal of Tianjin University Science and Technology*, 58(08). ISSN 0493-2137. https://doi.org/10.5281/zenodo.16948349

Kumar, S., et al. (2023). *Large-scale persuasion with language models*. arXiv:2307.12345.

*McCarron, M. (2024).* Battlespace of Mind

Mitchell, M., et al. (2023). *Fully autonomous AI agents should not be developed*. https://arxiv.org/html/2502.02649v3

Nasim, M., et al. (2025). *Simulating influence dynamics with LLM agents*. arXiv.

NATO StratCom COE. (2017–2023). *Handbooks on strategic communications and cognitive influence*.

Park, J. S., et al. (2023). *Generative agents: Interactive simulacra of human behavior*. https://arxiv.org/abs/2304.03442

Putra, R. (2023). *Autonomous systems and escalation control in military operations*. https://esaformosapublisher.org/index.php/esa/article/download/40/34

RAND Corporation. (2025). *Acquiring generative AI to improve DoD influence activities*. RAND.

Schmitt, P., & Flechais, I. (2024). *Digital deception: Generative AI in social engineering and phishing*. *Artificial Intelligence Review*. https://link.springer.com/

Singh, J. (2025). *Unleash(ed) AI: The rise of cognitive security operations*. https://blogs.cisco.com/customerexperience/unleashed-ai-the-rise-of-cognitive-security-operations

TechRadar. (2024). *AI-powered DDoS attacks and cybercrime trends*. https://www.techradar.com/pro/5-ways-ai-is-supercharging-ddos-attacks

Thomas, T. (2004). *Russia's reflexive control theory and the military*. *Journal of Slavic Military Studies*.

UN Office for Disarmament Affairs. (2023). *Algorithmic escalation and risks of flash warfare*. https://unric.org/en/ai-in-conflict-keeping-humanity-in-control/

UNODA. (2023). *Automated decision-making and algorithmic escalation: Risks of flash warfare*. United Nations.

World Economic Forum. (2024). *Global risks report 2024*. https://www.weforum.org/stories/2024/01/global-risk-report-2024-risks-are-growing-but-theres-hope/

Zhang, X., et al. (2025). *LLM-AIDSim: LLM agents for influence diffusion modeling*. *Systems*, MDPI. https://www.mdpi.com/2079-8954/13/1/29

Zhu, X., et al. (2025). *LLM-AIDSim: LLM agents for influence diffusion modeling*. *Systems*, MDPI.