The RUSI Journal

Routledge
Taylor & Francis Group

ISSN: 0307-1847 (Print) 1744-0378 (Online) Journal homepage: www.tandfonline.com/journals/rusi20

Characterising Cyber Cognitive Attacks

Bonnie Rushing & Shouhuai Xu

To cite this article: Bonnie Rushing & Shouhuai Xu (09 Mar 2026): Characterising Cyber Cognitive Attacks, The RUSI Journal, DOI: 10.1080/03071847.2026.2638130

To link to this article: https://doi.org/10.1080/03071847.2026.2638130

Published online: 09 Mar 2026.

Research Article
RUSI Journal

Characterising Cyber Cognitive Attacks

Bonnie Rushing and Shouhuai Xu

Cyber cognitive attacks are an emerging threat to society. However, our understanding of these attacks is superficial. This article aims to deepen understanding of these attacks by proposing a methodology to characterise them. Bonnie Rushing and Shouhuai Xu apply the methodology to conduct a case study based on 100 references. This leads to a novel definition of cognitive attacks, detection design, applications and requirements for effective defences against cognitive attacks.

Cyber cognitive attacks, or cognitive attacks for short, manipulate the subconscious thinking of targets by exploiting cognitive weaknesses, biases, mental shortcuts and reflexive thinking. These mechanisms shape an individual's immediate perception of reality[1] and can distort reasoning, influence decision-making and prompt hasty or irrational choices. Earlier influence activities, such as traditional propaganda, psychological operations (PSYOP) and information warfare, primarily sought to shape conscious belief formation through overt messaging and narrative control. Cognitive attacks extend these approaches by deliberately targeting subconscious cognitive processes that govern perception, judgement and action, often at scale with advanced technology and with increasing levels of personalisation.[2]

In this article, cognitive attacks are treated as discrete manifestations within the broader construct of cognitive warfare. Cognitive warfare should not be understood as a replacement for established doctrinal activities such as information warfare, PSYOP and military information support operations (MISO), but rather as an effects-level framework that explains the cognitive outcomes produced when these activities target human decision-making processes. Whereas doctrine such as the US Joint Publication 3-13 defines information warfare and MISO in terms of authorities, tools and functions for shaping the information environment, cognitive warfare describes the intended influence on perception, belief and choice that emerges from their strategic application.[3]

Accordingly, information warfare, PSYOP and MISO remain the doctrinal means of planning and execution, while cognitive warfare provides an integrating conceptual lens for understanding how and why these activities affect cognitive and narrative dynamics rather than merely transmitting information. This distinction directly supports the direction that the US Department of Defense defines cognitive warfare in relation to existing doctrine and assesses its alignment with information warfare, PSYOP and MISO, while examining narrative intelligence as a core activity within the cognitive domain.[4]

[1] Janna Anderson and Lee Rainie, 'The Future of Human Agency', Pew Research Center, 24 February 2023, <https://www.pewresearch.org/internet/2023/02/24/the-future-of-human-agency/>, accessed 21 January 2026.

[2] Bernard Claverie and François du Cluzel, '"Cognitive Warfare": The Advent of the Concept of "Cognitics" in the Field of Warfare', in Bernard Claverie et al., *Cognitive Warfare: The Future of Cognitive Dominance* (Neuilly: NATO-STO Collaboration Support Office, 2022), pp. 2-1–2-8.

[3] Joint Chiefs of Staff, 'Joint Publication 3-13: Information Operations – Incorporating Change 1', 20 November 2014.

[4] US Senate Committee on Armed Services, 'Narrative Intelligence and Cognitive Warfare' in US Senate Committee on Armed Services, 'National Defense Authorization Act for Fiscal Year 2026', Senate Report 119–39, 15 July 2025.

Cognitive attacks directly target humans' subconsciousness.
Courtesy of Oleksandr/Adobe Stock

Cognitive attacks are devastating for multiple reasons. From a technical point of view, cognitive attacks are becoming increasingly capable. They not only maliciously exploit advanced technologies such as artificial intelligence/machine learning (AI/ML), but also exploit democratic principles and processes.[5] For instance, cognitive attacks increasingly exploit the democratic principle of freedom of speech by spreading fake news or disinformation on social networks, silencing positive public debate through deception and weaponising public opinion and divisions.[6] From a non-technical point of view, cognitive attacks are devastating because they are allegedly often employed by nation-state actors to wage cognitive warfare. This appears to have become a new norm in conflicts below the threshold of conventional kinetic war. Thus, cognitive attacks are sowing 'distrust, skepticism, and instability'.[7]

Compared with other forms of cyberattacks, cognitive attacks are relatively new because they directly target humans' subconsciousness, arguably the least studied field from a cybersecurity perspective. As evidenced by the prevalence of disinformation in cyberspace, our understanding of these attacks remains limited, motivating the present study.

This article seeks to answer the following research questions: What are the defining characteristics of cognitive warfare in digital spaces? How should the term 'cyber cognitive attacks' be defined? How can the cybersecurity and social science communities better meet research gaps related to disinformation?

This article makes four contributions. First, it proposes a methodology for characterising cognitive attacks. Second, it demonstrates the methodology's usefulness by applying it to a case study based on 100 references. This leads to an innovative definition of cognitive attacks. To our knowledge, we are the first to aim for a definition that has the potential for broad adoption. Certainly, the definition could be seen as a good starting point for further improvement. Third, the article conducts a case study by applying the definition to analyse real-world cyber cognitive attacks, investigating the interplay of the web and society. The case study leads to several insights, including (i) cyber cognitive attacks have been evolving with increasing sophistication; and (ii) nation-state actors indeed appear to have been harnessing cognitive attacks with international security implications. Fourth, the article characterises the requirements of effective defences against

[5] Anderson and Lee, 'The Future of Human Agency'.
[6] François du Cluzel, 'Cognitive Warfare', NATO Innovation Hub/Allied Command Transformation, January 2021.
[7] Antony Blinken, speech given at the Summit for Democracy in Washington, DC, 9 December 2021.

DOI: 10.1080/03071847.2026.2638130

cognitive attacks to guide future research on designing effective defences. The requirements include building a knowledge base of cognitive attacks and creating a high-fidelity 'testbed' to test the efficacy of newly proposed defences.

The authors divide cognitive attacks into two categories: those occurring prior to the cyber era; and those emerging during the cyber era. In the period preceding the cyber era, strategists have long employed tactics designed to influence perception, decision-making and behaviour. Sun Tzu famously emphasised the centrality of deception, arguing that 'all warfare is based on deception'.[8] In *On War*, Carl von Clausewitz framed war as an extension of politics and described its objective as the disarming of the enemy, implicitly encompassing non-kinetic means alongside military force.[9] Similarly, in *The Prince*, Niccolò Machiavelli argued that effective rule often requires deception, noting that a leader must know how to manipulate appearances and mislead adversaries when necessary.[10] Deception has therefore been a fundamental element of what would now be described as cognitive attacks long before the advent of the cyber era.

In terms of cognitive attacks in the cyber era, malign actors have waged cyber cognitive attacks such as disinformation.[11] However, there is no widely accepted definition of cognitive attacks,[12] as evidenced by the presence of at least 100 explicit or implicit definitions explored in the current article. Alonso Bernal et al. investigated defining cognitive attacks as an emerging concept based on 27 prior studies.[13] Kathy Cao et al. analysed 56 corroborating sources while making a distinction between psychological attacks and cognitive attacks.[14] The cognitive domain often refers to individuals' or groups' minds and information processing, perception, judgement and decision-making.[15]

By contrast, this article proposes a methodology for characterising cognitive attacks. It demonstrates its usefulness with a new definition based on 100 prior studies and explores how to apply the definition to characterise cognitive attacks. Not least, cognitive attacks are treated as a subset of cyber social engineering attacks.[16] These are broader as they may exploit human factors that would be less relevant to disinformation (for example, greed).

Methodology

This article's proposed methodology is applied to guide the case study, but can also be used by other researchers to conduct further studies. Recognising that cognitive attacks lack a settled or universally accepted definition, the methodology is structured to systematically capture and evaluate the full range of relevant academic and policy literature. The methodology has the following steps: (1) defining attributes to describe cognitive attacks; (2) identifying cognitive attacks described in the literature, including both non-academic and academic sources; (3) applying the attributes to characterise the cognitive attacks described in the literature and draw insights; (4) leveraging (1), (2) and (3) to propose a new definition of cognitive attacks; (5) proposing requirements of desired solutions; and (6) identifying gaps and research directions.

Step 1 defines the attributes that can be adequately used to characterise cognitive attacks. The attributes should be defined based on domain knowledge, literature or both. These attributes may be qualitative or quantitative. Given that cognitive attacks are elusive to define, meaning that cognitive attacks may not have been explicitly defined in the literature, this article considers implicit discussions. Moreover, academic and non-academic literature must be considered as professionals may have widely discussed cognitive attacks.

Step 2 leverages the attributes identified in the preceding step to characterise cognitive attacks. Again, academic and non-academic literature containing implicit or explicit descriptions of cognitive attacks may be useful. This is achieved by keyword search and search engines, meaning

[8] Sun Tzu, *The Art of War*, translated by Samuel B Griffith (Oxford: Oxford University Press, 1963), Chapter 1.

[9] Carl von Clausewitz, *On War*, Book I, translated by Michael Howard and Peter Paret (Princeton, NJ: Princeton University Press, 1976), Chapter 1.

[10] Niccolò Machiavelli, *The Prince*, translated by Harvey C Mansfield (Chicago, IL: University of Chicago Press, 1998), Chapter XVIII.

[11] Andrew M Guess and Benjamin A Lyons, 'Misinformation, Disinformation, and Online Propaganda', in Nathaniel Persily and Joshua Tucker (eds), *Social Media and Democracy: The State of the Field, Prospects for Reform* (Cambridge: Cambridge University Press, 2020).

[12] Claverie and du Cluzel, '"Cognitive Warfare"', in Claverie et al., *Cognitive Warfare*.

[13] Alonso Bernal et al., 'Cognitive Warfare: An Attack on Truth and Thought', Johns Hopkins University, September 2020.

[14] Kathy Cao et al., 'Countering Cognitive Warfare: Awareness and Resilience', *NATO Review*, 20 May 2021.

[15] Joint Chiefs of Staff, 'Joint Publication 3-13'.

[16] Theodore Tangie Longtchi et al., 'Internet-Based Social Engineering Psychology, Attacks, and Defenses: A Survey', *Proceedings of the IEEE* (Vol. 112, Issue 3, 2024).

investigators must manually examine and filter the returned literature by relevancy.

Step 3 characterises the descriptions or definitions of cognitive attacks via the attributes. The resulting characteristics deepen our understanding and lead to valuable insights.

Step 4 leverages the characterised attributes to produce a definition of cognitive attacks that can accommodate as many attributes of cognitive attacks as applicable.

Steps 5 and 6 aim to identify knowledge and practitioner gaps in state-of-the-art cognitive attack research. The article proposes requirements and recommended projects to optimise and unify research efforts to develop understanding and defences further, seeking to bridge the gaps.

Case Study: Applying the Methodology to Cognitive Attacks

We demonstrate the value of the methodology by conducting the following case study.

Defining Attributes of Cognitive Attacks

Cognitive attacks target human subconsciousness by exploiting biases, mental shortcuts and/or reflexive thinking. This subconscious thought, 'System 1 thinking', is part of a person's immediate reaction and reality. It is where biases are embedded, forming deeply rooted credulity that provokes thought distortions, adversely influencing the victims' decision-making. Cognitive attacks impact victims' sensations, attention, memory and mental operations to mislead them into fallacious reasoning and manipulated decisions.[17] In contrast to System 1, 'System 2 thinking' impacts analytic or system processing. Attackers aim to cause targets to deviate from their original System 1 mental processes.[18]

Attribute Selection Methodology

A hundred published descriptions of cognitive warfare were gathered. These texts were recorded on a spreadsheet organised by their authors' articles. Next, the 100 descriptions were entered into ChatGPT 4o large language model (LLM). We searched for the top 10 trending attributes in the dataset to perform text analysis and attribute selection, prioritised by appearance frequency and their centrality to cognitive warfare. The LLM returned 10 attributes that our research team, as the human decision-makers, merged into seven attributes. Initially, the LLM listed these 10 attributes: perception manipulation; decision-making influence; information control/manipulation; psychological operations; technology enablers; population-level targeting; narrative warfare; behaviour modification; domain integration; and cognitive bias exploitation. Four suggestions from the LLM were selected: technology; information manipulation; population; and domains. Other ideas were combined for clarity. First, we combined these ideas about human cognitive exploitation: perception manipulation; psychological operations; narrative warfare; and cognitive bias exploitation. The attribute is named 'perception'. Further, we combined these ideas about influencing targets' actions: decision-making influence; and behaviour modification. The attribute is named 'action'. Finally, we combined the ideas for targeted warfare and PSYOP into the attribute 'weapon'.

These attributes are justified by general system theory and cybernetics: 'We cannot reduce the biological, behavioural, and social levels to the lowest level, that of the constructs and laws of physics'.[19] However, general system theory can be used to unify scientific disciplines and gain a more realistic understanding. The connections among various attributes were mapped as follows. Certain attributes are for interdependent system components from various fields including economics, politics, social, cyber and military (attribute: domains).

[17] François du Cluzel (ed), 'Cognitive Warfare'.

[18] Theodore Tangie Longtchi et al., 'Internet-Based Social Engineering Psychology, Attacks, and Defenses'.

[19] Ludwig von Bertalanffy, *General System Theory: Foundations, Development, Applications* (New York, NY: George Braziller, 1968).

Cybernetic feedback schemes are widely used in technology (tech) and human communications when the effector responds to message stimuli. A human target's perception (perception) may be influenced, and, subsequently, their behaviours may be triggered by these systems (action). Ludwig von Bertalanffy compares this complex system to George Orwell's *1984*, including the evolution of technology's destructive nature (weapon), methods of mass suggestion (population) that may affect larger communities, and examines human morals, logic and manipulation (info manipulation).[20]

This methodology prompts us to propose characterising cognitive attacks from the following seven attributes:

- Information manipulation: This attribute describes whether a cognitive attack manipulates the information presented to a target.
- Perception: This attribute describes whether a cognitive attack affects a victim's perceptions and thoughts.
- Action: This attribute describes whether a cognitive attack impacts a victim's decisions and actions.
- Population: This attribute describes whether a cognitive attack targets large groups such as communities, societies, militaries, countries or world populations.
- Tech: This attribute indicates whether a cognitive attack exploits modern technologies such as AI/ML to amplify its success.
- Weapon: This attribute indicates whether a cognitive attack is used intentionally as a weapon or attack against targets.
- Domains: This attribute describes whether a cognitive attack is executed concurrently with other operations in other domains (for example, alongside air or land operations in a military context) or national instruments of power (for example, economic or diplomatic actions).

The relationship between the attributes can be understood as follows. The first three attributes formulate the core of cognitive attacks as information manipulation represents a primary means used by a cognitive attacker to affect a target's perception and thoughts, and thus impact the target's decision-making or actions. The population attribute describes the scale of a cognitive attack, which may be amplified by leveraging modern technology to achieve the scale. The last two attributes are relevant in military contexts, namely that cognitive attacks may be used as weapons in the cognitive domain, possibly concurrent to operations in the other domains of conflicts.

Systematic Literature Search and Screening Process

To identify the relevant description of cognitive attacks in published literature, the search strategy included Google and Google Scholar. It discovered literature via the following Boolean query:

> cognitive AND (warfare OR operations OR attack)

This syntax was also used to search the Air University Library catalogue. We screened literature by titles and abstracts to filter relevant studies and applied the inclusion and exclusion criteria to the full texts.

The included pieces consisted of: articles that discuss cognitive attacks, disinformation or influence operations in digital spaces; high-quality reports or sources such as peer-reviewed journal articles, conference papers or government programmes; and papers published in English (or translated works).

The excluded pieces consisted of: non-academic sources such as personal blogs, message boards or opinion pieces; articles that do not focus on cognitive attacks or their attributes; and duplicate studies without substantial contributions.

In filtering the search results, academic journals were prioritised. Most literature provides implicit descriptions of cognitive attacks, few provide explicit definitions. Nevertheless, we observed non-peer-reviewed articles that also discussed cognitive attacks. This prompted us to include some of them due to their quality. In total, this led to 100 references.

[20] *Ibid.*

Applying the Attributes to Characterise Literature Descriptions

We highlighted and hand-coded the previously mentioned reoccurring trends (attributes) within the 100 authors' presentations of their most succinct definition of 'cognitive attack'. If a publication included other trending criteria elsewhere in writing, outside of the closest succinct description (typically one to two sentences) of cognitive attacks, it was not included or coded for this definition analysis. The following steps were taken to categorise the seven trending attributes:

1. Examination of 100 definitions from published literature, using well-established cognitive and behavioural science theories and ChatGPT - 4o LLM analysis to assist in discovering frequent trends. We call these frequent reoccurrences 'attributes'.
2. Rescrutinising the definitions and highlighting every applicable attribute in each publication's writing. We categorised the highlighted rhetoric based on the overt use of attributes or synonymous terminology.[21]

Examples from our manual coding of attribute terminology are as follows:

- Info manipulation coding examples: manipulating environmental stimuli; manipulation of the public discourse; disinformation; disseminate misleading narratives; coordinated inauthentic behaviour; algorithmic manipulation; and deepfakes.
- Perception coding examples: modify perceptions of reality; alter the cognition of human targets; create a skewed sense of reality, subconscious; bypass our rational conscious mind; and exploit biases, fallacies, emotions and automatisms.
- Action coding examples: behaviours; decisions and actions; hinder action; how the target population behaves; responded to or acted on; and decision-making process.

Population coding examples: population-level cognition; whole-of-society manipulation; to a large extent, nations; undermine social unity; and communities.

Tech coding examples: using technological tools, advanced technologies to target; and employing digital technologies, social media platforms, and nanotechnology, biotechnology and information technology.

Weapon coding examples: critical realm of warfare; gain an advantage over an adversary; weaponisation of public opinion; the realm of neuro-weapons; direct combat targets; and persuasion wars.

Domains coding examples: synchronisation with other instruments of power; with the aid of additional warfare tools.

Characterising Concrete Cyber Cognitive Attacks

Thirty eight of the 100 analysed sources included the 'tech' attribute. Information technology is thus clearly linked to modern cognitive attacks and will continue to compound attackers' dangerous effects. The study applied technologies and the defined attributes to the following real-world example of a cyber cognitive attack: the 2016 US presidential election campaign, when hundreds or thousands of Russian fake accounts posted anti-Clinton messages, such as 'Hillary was a criminal'.[22]

The effects of this example's cognitive attack may have led victims to reconsider their presidential candidate preferences or caused other cognitive impacts such as societal polarisation, fomenting political strife or skewing online discourse.[23] To achieve these objectives, Russian actors may have operated as indicated below.

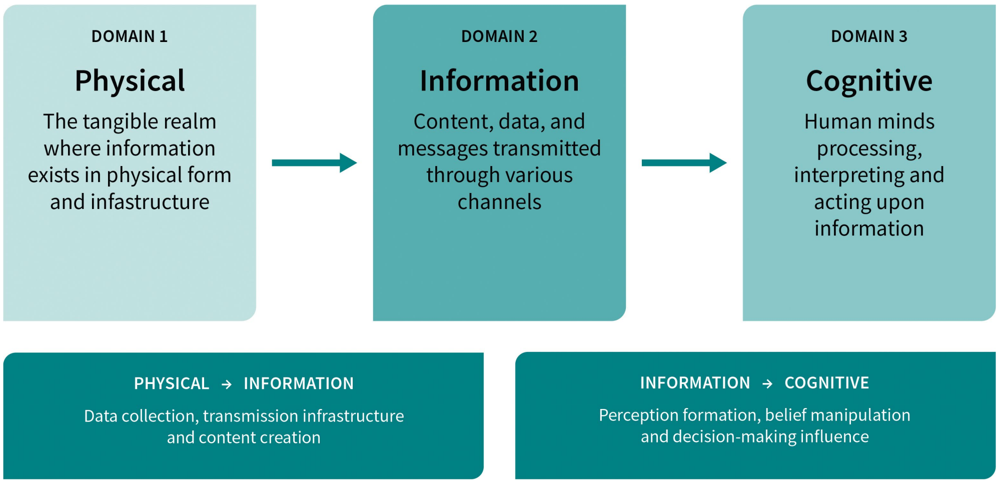
Russian actors may conduct cyber cognitive attacks in conjunction with multi-domain warfare (attribute: domains). Russian actors may communicate with targets using modern technologies (tech) such as social media, email,

[21] The results are available online. See Bonnie Rushing, 'Cognitive Attack Attribute Comparison Chart', <https://thebonnierushing.com/defining-cognitive-attack>, accessed 21 January 2026.

[22] Hunt Allcott and Matthew Gentzkow, 'Social Media and Fake News in the 2016 Election', *Journal of Economic Perspectives* (Vol. 31, No. 2, Spring 2017).

[23] Stefan Wojcik et al., 'Bots in the Twittersphere', Pew Research Center, 9 April 2018, <https://www.pewresearch.org/internet/2018/04/09/bots-in-the-twittersphere/>, accessed 21 January 2026.

Figure 1: Information Environment Dimensions


DOMAIN 1
Physical
The tangible realm where information exists in physical form and infastructure
DOMAIN 2
Information
Content, data, and messages transmitted through various channels
DOMAIN 3
Cognitive
Human minds processing, interpreting and acting upon information
PHYSICAL → INFORMATION
Data collection, transmission infrastructure and content creation
INFORMATION → COGNITIVE
Perception formation, belief manipulation and decision-making influence

Source: Author generated. Adapted from Joint Chiefs of Staff, 'Joint Publication 3-13: Information Operations – Incorporating Change 1', 20 November 2014.

direct messaging, gaming platforms, virtual reality devices and many apps on Internet of Things devices,[24] using content in the information domain, delivered to the physical domain, to affect the cognitive domain, as depicted in Figure 1. These hundreds or thousands of fake accounts may be referred to as a 'troll factory',[25] and their targeting and operations are enhanced by tracking and data collection of users' data and preferences,[26] including data of individuals or broader populations (population). They employ bots to model human communication behaviour[27] and weaponise (weapon) algorithmic targeting to customise content for targeted groups. They may also intentionally spread false messaging (info manipulation), including disinformation and AI-assisted synthetic media such as deepfakes.[28] Attackers may use LLMs to generate realistic, persuasive and tailored cognitive attack content[29] to modify perceptions of reality (perception) and perhaps influence targets' decisions and behaviours (action).[30]

These emerging technologies aggravate an escalatory arms race between cyberspace users and cognitive attackers. These technologies amplify the reach and deepen the impacts of cognitive attacks. Thus, defensive measures must also be amplified.

[24] Michael Bossetta, 'The Digital Architectures of Social Media: Comparing Political Campaigning on Facebook, Twitter, Instagram, and Snapchat in the 2016 U.S. Election', *Journalism & Mass Communication Quarterly* (Vol. 95, No. 2, 2018).

[25] Adrian Chen, 'The Agency', *New York Times Magazine*, 2 June 2015, <https://www.nytimes.com/2015/06/07/magazine/the-agency.html>, accessed 21 January 2026.

[26] Joseph Turow, Michael Hennessy and Nora A Draper, 'The Tradeoff Fallacy: How Marketers are Misrepresenting American Consumers and Opening Them Up to Exploitation', University of Pennsylvania, June 2015.

[27] Wojcik et al., 'Bots in the Twittersphere'.

[28] Robert Chesney and Danielle K Citron, 'Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security', *California Law Review* (Vol. 107, No. 6, 2019).

[29] Josh A Goldstein et al., 'Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations', arXiv:2301.04246, January 2023; du Cluzel (ed), 'Cognitive Warfare'; James Johnson, 'Artificial Intelligence: A Threat to Strategic Stability', *Strategic Studies Quarterly* (Vol. 14, No. 1, Spring 2020).

[30] Josh A Goldstein et al., 'Generative Language Models and Automated Influence Operations'.

New Definition of Cognitive Attacks

Using the study's methodology, we evaluated and systematised 100 cognitive attack definitions by scrutinising their terminology and highlighting the popular ideas (attributes) used to describe cognitive attacks. Next, we categorised each definition according to the trends evident in its rhetoric. We developed a comparison chart listing each publication in a row and marking each column that applies to the written definition of the author(s). We tracked these trends in the comparison chart's columns: information is modified or manipulated (info manipulation); affects perception and thoughts (perception); impacts targets' decision-making or actions (action); impacts large populations (population); described as a weapon or attack (weapon); technology-focused (tech); and employed with other warfighting domains (domains). Finally, we added the trend column totals to discover the percentage of authors who included the trending ideas in their definitions.

For example, one publication's description states, 'Cognitive warfare is the art of using technology to alter the cognition of human targets, who are often unaware of any such attempt'.[31] This definition indicates two trends: impacts on perceptions or thoughts; and a technology focus. We organised these trends in a cognitive attack comparison chart across the X-axis. The sources are cited on the Y-axis in no particular order, with 'x' marks in each corresponding block where each trend was found in each publication. Our team included the attribute content and statistics to form an updated and standardised definition of 'cognitive attack' based on the 100 published descriptions. At its core, cognitive attack refers to operations that target human minds to gain advantages. We developed this core understanding with details based on statistical terminology reoccurrences. Our methodology for defining cognitive attacks is based on occurrence frequency. For example, the trending idea(s) are included as a concrete part of the definition only if 70–100% of the analysed publication descriptions contain it.

To include trends in our novel definition, we use the following statistical methodology, applied in Table 1:

- 70–100% = concrete part of the definition.
- 50–69% = 'typical' in definition.
- 30–49% = 'may/possible' in definition.
- Below 30% = not included in the definition.

Table 1: Statistics of Cognitive Attack Descriptions

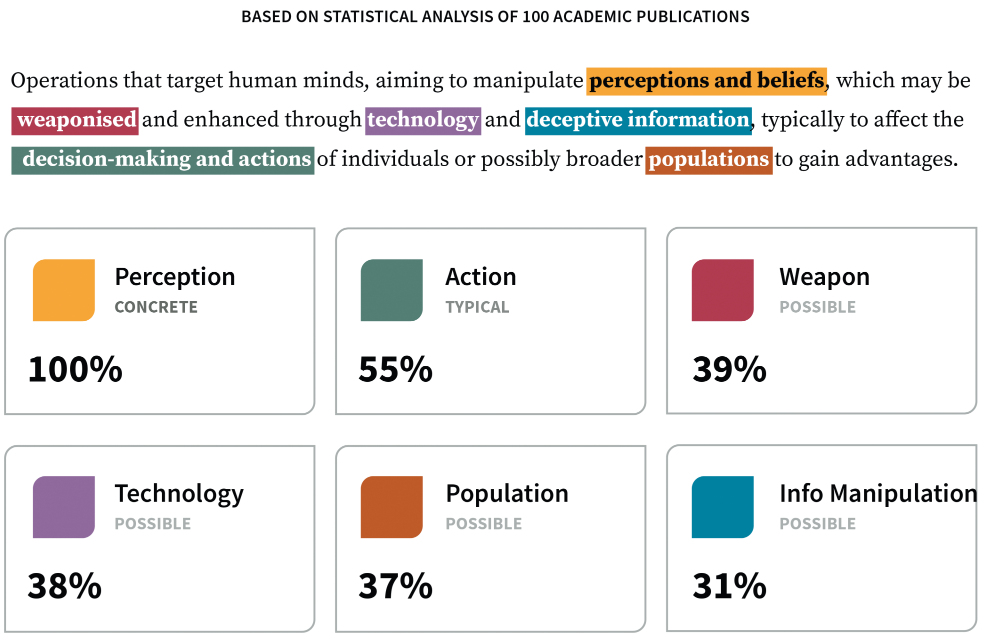
Trend	Reoccurrence	Our Verbiage
Info manipulation	31%	May/possible
Perception	100%	Concrete
Action	55%	Typical
Population	37%	May/possible
Weapon	39%	May/possible
Tech	38%	May/possible
Domains	10%	Not included

Source: The authors.

> Attackers may use LLMs to generate realistic, persuasive and tailored cognitive attack contentto modify perceptions of reality (perception) and perhaps influence targets' decisions and behaviours (action).

[31] Claverie and du Cluzel, 'Cognitive Warfare', in Claverie et al. (eds), *Cognitive Warfare*.

Figure 2: Cognitive Attacks Definition



Operations that target human minds, aiming to manipulate **perceptions and beliefs**, which may be **weaponised** and enhanced through **technology** and **deceptive information**, typically to affect the **decision-making and actions** of individuals or possibly broader **populations** to gain advantages.

Source: The authors.

Defining Cognitive Attacks

From here, we propose the following standardised 'cognitive attack' definition: operations that target human minds, aiming to manipulate perceptions and beliefs, which may be weaponised and enhanced through technology and deceptive information, typically to affect the decision-making and actions of individuals or possibly broader populations to gain advantages.

This definition incorporates statistically prevalent ideas highlighted in Figure 2, which shows how each attribute is included in the definition's verbiage. As this is a concise definition, it does not include all details of the threat. Specifically, cognitive attacks target the subconscious, exploiting mental biases and shortcuts. This subconscious thought, known as System 1 thinking, becomes part of a person's immediate reaction and reality.[32] Cognitive attacks reshape individuals' deep underlying knowledge to manipulate their subconscious perceptions and reactions.[33] Further, they impact targets' sensation, attention, memory and mental operations. Victims are misled into fallacious reasoning and manipulated decisions.[34]

This article's findings corroborate established foundational research in cognitive psychology and information influence. The results demonstrate that cognitive attacks primarily target the human mind, with the central objective of influencing subconscious cognitive processes. All analysed articles (100%) indicate that cognitive attacks affect targets' perception and thought processes, while 55% further suggest that altered perception translates into changes in decision-making and behaviour. In addition, cognitive attacks

[32] Anderson and Lee, 'The Future of Human Agency'.

[33] Yuriy Gorodnichenko, Tho Pham and Oleksandr Talavera, 'Social Media, Sentiment and Public Opinions: Evidence from #Brexit and #USElection', NBER Working Paper No. 24631, May 2018.

[34] Elizabeth F Loftus, 'Planting Misinformation in the Human Mind: A 30-Year Investigation of the Malleability of Memory', *Learning & Memory* (Vol. 12, No. 4, 2005).

frequently occur at scale: 37% of the reviewed definitions explicitly identify groups or entire societal populations as targets. Moreover, 10% of the sources indicate that cognitive attacks occur concurrently with activities in other domains, including economic and political action. These combined efforts frequently transcend sovereign borders, targeting the cognitive foundations of societies rather than isolated individuals.

These insights may aid interdisciplinary researchers as cognitive attack studies continue. Further, practitioners and policymakers may incorporate these data into updated doctrine, policy, defensive operations or threat modelling.

Application

Empirical Analysis

Strongly Correlated Attribute Pairs

The study analysed the correlations between attributes frequently found within a reference, discovered strongly correlated pairs, and developed the following quantitative and qualitative insights:

- Perception and action (co-occurrence in 55 references).
  – Insight: Cognitive attacks often focus on altering perception to drive specific actions.
- Perception and info manipulation (co-occurrence in 31 references).
  – Insight: Suggests that manipulation of information plays a critical role in altering perceptions.
- Perception and population (co-occurrence in 37 references).
  – Insight: Indicates a frequent focus on large-scale perception impacts.
- Perception and tech (co-occurrence in 38 references).
  – Insight: Highlights the role of technology in amplifying cognitive attacks targeting perception.

Cognitive Attack and Defence Simulation

Our team simulated cognitive attacks in a military wargame with original fictional media communications. The immersive content included simulated news, radio broadcasts and social media posts applicable to the scenario. We created a fictional social media post example that may be used in a cognitive attack simulation, depicted in Figure 3. About 87% of participants indicated that these media influenced their decision-making in the wargame.[35] Simulated media environments (such as the conducted wargame) benefit from incorporating our cognitive attack attributes. Specifically, the methodology was applied by creating realistic fictional media that was presented to participants throughout the scenario's timeline. The wargame's information environment applied the attributes as follows:

1. Info manipulation: Simulated disinformation and misleading narratives were broadcast.
2. Perception: Media aimed to skew targets' sense of reality and impact emotions.
3. Action: Media aimed to impact the decisions and actions of the audience.
4. Population: Media targeted whole communities or nations.
5. Tech: Media was posted through technological tools such as social media platforms.
6. Weapon: Some communications were broadcast by a simulated adversary seeking to gain a strategic advantage.
7. Domains: Information environment simulated alongside multi-domain operations.

Including these attributes improved the realism of cognitive attack simulation for participants and researchers.

> Specifically, the methodology was applied by creating realistic fictional media that was presented to participants throughout the scenario's timeline.

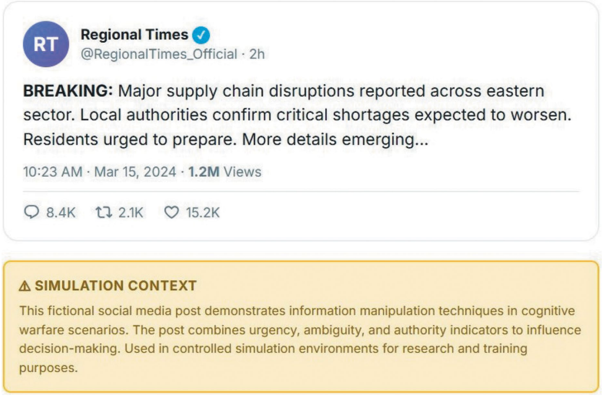
[35] Bonnie Rushing, 'Analysis of Media Influence on Military Decision-Making', in Jaco du Toit and Prof Brett van Niekerk (eds), *Proceedings of the 19th International Conference on Cyber Warfare and Security* (Vol. 19, No. 1, 2024).

Furthermore, media literacy education was tested as a defence, and its effectiveness against these simulated cyber cognitive attacks was evaluated. Data revealed that wargame participants who completed a media literacy lesson module:

- Thought about media 32.6% more often.
- Discussed media with peers 44% more often.
- Considered source trustworthiness 34.9% more often.
- Were 12% more likely to trust unbiased/data-driven sources.
- Were 23% more suspicious of disreputable sources.[36]

These findings show the possible benefits of media literacy education in combating disinformation. This methodology can be employed to test other defensive measures, populations, scenarios and simulation variables.

Figure 3: Simulated Social Media Post
Example for Cognitive Warfare Simulation



Regional Times
@RegionalTimes_Official · 2h

BREAKING: Major supply chain disruptions reported across eastern sector. Local authorities confirm critical shortages expected to worsen. Residents urged to prepare. More details emerging...

10:23 AM · Mar 15, 2024 · **1.2M** Views

8.4K 2.1K 15.2K

⚠ SIMULATION CONTEXT

This fictional social media post demonstrates information manipulation techniques in cognitive warfare scenarios. The post combines urgency, ambiguity, and authority indicators to influence decision-making. Used in controlled simulation environments for research and training purposes.

Source: The authors.

Explainable AI for Cognitive Attack Detection

Our framework can automatically be used to build online tools to detect cyber cognitive attack content. Developers should prioritise transparency in these automated detection systems by explaining why specific content was flagged as cognitive attack material. For example, 'This post was flagged due to the presence of information manipulation and intent to influence decision-making tactics'. These transparent interpretation techniques will highlight our defined attributes and build trust with users. Using our methodology, users will learn about cyber cognitive attack attributes, security threats and applications.

Empirical Real-World Framework Validation

We applied our definition framework for cyber cognitive attacks to recent real-world case studies for validation. This methodology helps investigators to better understand real-world events using our new definition and the seven attributes. Example attacks are depicted in Table 2 and assessed as follows.

1. Cambridge Analytica Scandal 2016.[37] Six attributes were present. These included: info manipulation – manipulated data for tailored messaging; perception – shaped voter perception of political candidates; action – influenced voting behaviour; population–targeted large segments of the population; technology – relied on data analytics and social media; and domains – focused on political and social domains. The missing attribute: weapon (arguably not positioned as a weaponised attack).
2. Russian Disinformation Campaign – 2016 US Election.[38] Seven attributes were present: info manipulation – disinformation spread through fake news; perception – altered public perception of political realities; action – influenced voter decisions; population – operated on a societal scale; tech – utilised social media algorithms and bots; weapon – used as a tool of psychological warfare; and domains – political, social and cyber domains.
3. QAnon Conspiracy Theory.[39] Seven attributes were present: info manipulation – spread conspiracy theories/disinformation; perception – skewed individuals' perceptions of reality; action – led to real-world actions such as protests and riots; population – engaged large, diverse online communities; tech – amplified by social media algorithms/platforms; weapon – weaponised narratives to undermine trust in institutions; and

[36] *Ibid.*
[37] Nicolas Bontridder and Yves Poullet, 'The Role of Artificial Intelligence in Disinformation', *Data & Policy* (Vol. 3, 2021).
[38] Peter Pomerantsev, *This Is Not Propaganda: Adventures in the War Against Reality* (London: Faber & Faber, 2019).
[39] Bontridder and Poullet, 'The Role of Artificial Intelligence in Disinformation'.

Table 2: Cyber Cognitive Attack Case Studies and Attributes

Case study	Info manipu-lation	Perception	Action	Population	Tech	Weapon	Domains
Cambridge Analytica Scandal (2016)	X	X	X	X	X		X
Russian Disinformation Campaigns (2016 US Election)	X	X	X	X	X	X	X
QAnon Conspiracy Theory	X	X	X	X	X	X	X
Covid-19 Misinformation	X	X	X	X	X		X
Deepfake media	X	X	X	X	X	X	X
Stuxnet Worm	X	X	X	X	X	X	X
Hybrid Warfare (Ukraine Conflict)	X	X	X	X	X	X	X

Source: The authors.

domains – operated across social, political and online platforms.

4. Covid-19 misinformation.[40] Six attributes were present: info manipulation – false claims about treatments and vaccines; perception – altered public understanding of health risks; action – influenced vaccine hesitancy and public health behaviours; population – affected global populations; tech – spread through social media/messaging platforms; and domains – operated in health, social and online spaces. The missing attribute: weapon (although harmful, it was not consistently framed as an intentionally weaponised attack).
5. Deepfake media.[41] Seven attributes were present: info manipulation – altered content to spread false narratives; perception – created false perceptions of public figures; action – could influence actions based on fake content; population – potentially targets large audiences; tech – utilises AI and machine learning; weapon – can be weaponised to incite conflict; and domains – operates across media, political and social domains.
6. Stuxnet worm.[42] Seven attributes were present: info manipulation – manipulated data within nuclear systems; perception – created uncertainty about system reliability; action – delayed Iran's nuclear progress; population – indirectly impacted the population via geopolitical effects; tech – highly sophisticated malware; weapon – explicitly designed as a cyber weapon; and domains – operated in cyber and physical domains.
7. Hybrid warfare – Ukraine conflict.[43] Seven attributes were present: info manipulation – disinformation campaigns targeting Ukraine and international audiences; perception – skewed narratives about the conflict; action – influenced actions of citizens and governments; population – targeted large international populations; tech – utilised cyber tools and social media; weapon – a tool of conflict; and domains – cyber, military and political domains.

[40] Jon Roozenbeek et al., 'Susceptibility to Misinformation About COVID-19 Around the World', *Royal Society Open Science* (Vol. 7, No. 10, October 2020).

[41] Giorgio Patrini, 'Mapping the Deepfake Landscape', Sensity AI, 7 October 2019, <https://giorgiop.github.io/posts/2018/03/17/mapping-the-deepfake-landscape/>, accessed 21 January 2026.

[42] Ralph Langner, 'Stuxnet: Dissecting a Cyberwarfare Weapon', *IEEE Security & Privacy* (Vol. 9, No. 3, May 2011).

[43] Anna Romandash, 'Hybrid Warfare: Ukraine, Russia and Western Lessons', Centre for International Governance Innovation, Policy Brief No. 209, 29 September 2025, <https://www.cigionline.org/publications/hybrid-warfare-ukraine-russia-and-western-lessons/>, accessed 21 January 2026.

These cases demonstrate the applicability of our methodology and the nature of real-world cyber cognitive attacks.

Societal Applications

The methodology can influence policymaking and societal applications in combating cognitive attacks. Examples of applications include:

- discover and determine the cyber cognitive attack status of events, content and case studies;
- expose malicious actors and reduce their effects through a clear understanding of cognitive attack attributes;
- strengthen trust in legitimate media sources through transparency and explainable AI;
- highlight tactics and disinformation noise and explain and diminish attackers and their motivations;
- run simulations with cyber cognitive attack attributes to design defences and education.

Requirements to Fill Gaps

This field has knowledge and practitioner gaps, a lack of shared understanding of cognitive attacks, and a lack of unified lines of effort. Based on our discoveries from developing a cognitive attack definition and analysing real-world cases, we present characterised requirements that can guide the design of future research and solutions to defeat these attacks. Cyber cognitive attack studies need formal requirements to optimise efforts and develop defences. We propose the following research directions with example projects based on completed research, continuously emerging technology and escalating security threats.

> Cyber cognitive attack studies need formal requirements to optimise efforts and develop defences.

Technology

As technology and tactics evolve, researchers must maintain pace and relevancy with the cyber cognitive attack threat model. It could be used to expand frameworks to employ emerging technologies such as generative AI. Our attribute labels could be used to detect non-text content in future experiments with natural language processing. This could be used to analyse actions in media to predict if case studies fall under the cognitive attack category.

Offensive

Adversaries, including government actors, employ offensive cyber cognitive attacks. These secretive actors are unlikely to share their playbooks, so researchers must discover and quantify the effects of their techniques. Example projects include quantifying cognitive attacks' sophistication, effectiveness and human susceptibility. It could also integrate cognitive attack frameworks with machine learning models for predictive analysis.

Defence

Researchers must continue to test defence efficacy and discover new solutions to mitigate cognitive attack threats. Current defence efforts include digital literacy skills and other educational programmes, internet policy controls, cyber detection and the TikTok ban debate.[44] Example projects include designing countermeasures to thwart cognitive attacks. It could also integrate automated disinformation and deepfake detection analysis.

Broad Field Inclusion

This includes exploring how cognitive attacks interact with other domains (for example, financial, diplomatic) to analyse cross-domain impacts. It could also integrate advanced technical education on influence operations, information integrity and social computing, connecting interdisciplinary projects.

[44] Gorodnichenko, Pham and Talavera, 'Social Media, Sentiment and Public Opinions'.

Discussion

International Security Implications

Individuals, non-state actors, states and international government organisations can conduct cognitive operations. Attacks take on many forms, with varied motivations and goals. Some focus on reinforcing groups' or individuals' existing ideals, while other operations seek to disrupt cohesion or accepted beliefs (perception),[45] seeking to influence the decisions of victims (action). Cognitive warfare is a significant and ever-evolving global conflict and security issue.[46] Cognitive conflict does not provide attackers with instantaneous rewards or success. It is typically a long process with multiple channels of propagating the selected narrative in print, television, social media, academia and political discourse.[47] US Secretary of State Antony Blinken warned that hostile cognitive attack activities are sowing 'distrust, skepticism, and instability' in democracies worldwide.[48] Cognitive operations exploit democratic principles such as freedom of speech especially as technology expands. Social networks enable opportunities to spread 'fake news', part of disinformation (info manipulation), silence public debate through deception and weaponise public opinion and divisions (weapon).[49]

At state levels, some leaders prioritise cognitive operations, using them alongside other instruments of power (domains). For example, Russia's campaign to influence the system, primarily led by its Internet Research Agency, uses social media platforms such as Facebook, Twitter/X and YouTube (tech), as well as Kremlin-associated media outlets. These operations fall under Russia's umbrella of active measures.[50]

> The growing threat of cognitive attacks threatens populations worldwide, and yet, few defensive measures have been standardised and implemented.

Additionally, China and the People's Liberation Army (PLA) consider cognitive attacks equal to other domains such as air, sea and space. They believe it is vital to victory, even without kinetic war. They exploit the openness of democratic societies' information infrastructure.[51] As part of this strategy, the PLA executes 'public opinion warfare' (population) operations.[52] This priority is integrated into Beijing's military strategy, with the 'intelligent warfare' concept that appears in its 2019 defence white paper.[53] China considers the cognitive domain a component of modern warfare and the centre of gravity that decides the outcome of war.[54] Chinese writers emphasise the importance of data, algorithms and computing power to intelligent warfare. They stress that data is the foundation of intelligent warfare, claiming that data is the 'new oil' and big data is the 'most important resource'. The importance of algorithms in intelligent warfare is related to the central role of AI. Chinese writers claim that warfare will become a contest between competing algorithms and that real-time computing of large amounts of data is essential (tech).[55] These cognitive attacks threaten global populations.

The growing threat of cognitive attacks threatens populations worldwide, and yet, few defensive measures have been standardised and implemented. Examples of cognitive attack defences include education programmes such as digital literacy and internet policy controls.

[45] François du Cluzel, 'Cognitive Warfare: A Battle for the Brain', NATO Science and Technology Organization, STO-MP-AVT-211-KN3, 2020.

[46] Tzu-Chieh Hung and Tzu-Wei Hung, 'How China's Cognitive Warfare Works: A Frontline Perspective of Taiwan's Anti-Disinformation Wars', *Journal of Global Security Studies* (Vol. 7, No. 4, 2022).

[47] Anderson and Lee, 'The Future of Human Agency'.

[48] Blinken, speech given at the Summit for Democracy.

[49] Romandash, 'Hybrid Warfare: Ukraine, Russia and Western Lessons'; Francesco Giumelli, 'The Redistributive Impact of Restrictive Measures on EU Members: Winners and Losers from Imposing Sanctions on Russia', *Journal of Common Market Studies* (Vol. 55, No. 5, 2016).

[50] Guess and Lyons, 'Misinformation, Disinformation, and Online Propaganda'; *ibid*.

[51] Gorodnichenko, Pham and Talavera, 'Social Media, Sentiment and Public Opinions'.

[52] Elsa Kania, 'The PLA's Latest Strategic Thinking on the Three Warfares', *China Brief* (Vol. 16, No. 13, 2016); Nathan Beauchamp-Mustafaga, 'Cognitive Domain Operations: The PLA's New Holistic Concept for Influence Operations', *China Brief* (Vol. 19, No. 16, 2019).

[53] People's Republic of China, *China's National Defense in the New Era* (Beijing: Foreign Languages Press, 2019).

[54] Hung and Hung, 'How China's Cognitive Warfare Works'.

[55] Elsa B Kania and John K Costello, 'The Strategic Support Force and the Future of Chinese Information Operations', *Cyber Defense Review* (Vol. 3, No. 1, Spring 2018).

The cyber-reliant public becomes increasingly vulnerable to these attacks as private information is compromised online.[56] Humanity must optimise efforts to develop defences for these dangerous cognitive attacks.

Limitations

The present study has the following limitations. The literature cited is in English or translated into English, meaning the findings may not necessarily align with descriptions published in other languages. We attempted to mitigate this narrow scope with our multi-lingual research team. Further investigation may benefit from leveraging machine learning to replicate our manual hand-labeling process, improving methodology subjectivity. Furthermore, this study considers attribute frequency as the key to cognitive attack definition, but it may also be relevant to consider the word position, token sequence and context.

Ethical Considerations

We address potential ethical considerations as follows. This research on cognitive attacks is not meant to aid malign actors. It is theoretical, and we disapprove of cyber cognitive attack offences. Since malicious actors may evade our attribute search terminology and labels, we recommend continuously updating definitions and expanding capabilities by advancing natural language processing to discover online disinformation content beyond our example text labels' rhetoric. Finally, technologies such as LLMs must be combined with human expertise when analysing content such as cognitive warfare definitions. The authors led our team's final data interpretation and determinations, ensuring the results aligned with community standards.

Conclusion

This article has presented a methodology to characterise cognitive attacks. It has also applied the methodology to conduct a case study based on 100 references, resulting in a novel definition of cognitive attacks and several insights. Both the definition and the requirements that are proposed for effective defences against cognitive attacks are subject to further refinement. Still, this approach can guide future studies on understanding and defending against cognitive attacks. ■

Bonnie Rushing is Senior Master Sergeant in the US Air Force and a PhD researcher in Security at the University of Colorado Colorado Springs.

Shouhuai Xu is Gallogly Chair Professor in Cybersecurity, Department of Computer Science, University of Colorado Colorado Springs.

The views expressed are those of the authors and do not reflect the official policy or position of the US Air Force Academy, US Air Force, Department of Defense or the US government.

[56] Roberto Baldoni and Giuseppe Di Luna, 'Sovereignty in the Digital Era: the Quest for Continuous Access to Dependable Technological Capabilities', arXiv: 2503.10140, 13 March 2025.