Cognitive Security Institute

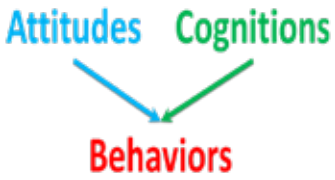
Dr. Matthew Canham
Executive Director, Cognitive Security Institute
Matthew@CognitiveSecurity.Institute

Mike Elkins
CH&ISO, BAMFIST and Humanis Technologies

What's in Your Cognitive Security Plan?
Human and AI Risk Planning

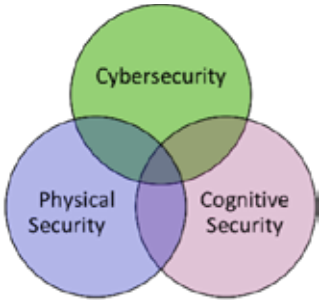
RSAC 2025 — [LAB2-T09]

Objectives of a Cognitive Attack



The objectives of cognitive attacks are to induce or influence a behavioral outcome or action. For example, an attacker employing a business identity compromise attack (employing either email, deepfake impersonation, or another method) uses deception to manipulate the target's cognitive state to lower suspicions and convince them that transferring money to a desired account is authorized (when it is not)[1]. Similarly, an influence operation may employ persuasion to manipulate a target's attitude toward their own company with the goal of convincing that target to commit an act of sabotage against their employer[2]. Regardless of the form that a cognitive attack takes (phishing, propaganda, LLM jailbreaking) an attack will always follow a similar pattern of manipulating attitudes or cognitions in order to influence or induce a desired behavioral outcome.

Multi-Domain Security



As the world becomes increasingly integrated, security professionals need to become acutely sensitive to the integration of security domains through the inter-relationship of cyber, physical, and cognitive systems.

- **Physical Security** focuses on securing physical assets and preserving the safety of individuals.
- **Cyber Security** is concerned with securing informational systems and assets.
- **Cognitive Security** by necessity focuses on preserving the integrity of decision-making mechanisms, primarily in humans but also in other information processing systems such as AI (artificial intelligence).

Attackers may leverage the inter-relation between domains and exploit gaps in security systems because of bureaucratic siloing. ICE (induce covert effect) attacks present an example of how threat actors may deliberately exploit these gaps by intentionally manipulating factors in one domain, to cause an action in a second (intermediate) domain, to induce an effect in a third domain (the target). For example, there have been multiple documented examples of armed robbers or child predators placing beacons in the game layer of PokemonGo (cyber domain) to entice their victims to view the character in augmented space (cognitive domain) which motivates the victim to move to the geo-correlated physical location where

they may be victimized (physical domain).[3,4,5] This style of attack may be used by any threat actor motivated to cause harm to a specific individual by either modifying their course of travel by inducing virtual vehicle traffic congestion[6,7], or placing virtual objects in geo-co-located positions.

Another proof-of-concept ICE attack employed online discount offers (cyber domain) for using electricity during specific times of the day, which motivated people to take advantage of those offers (cognitive domain)—even when those discount usage times corresponded with peak power usage and would cause grid collapse (physical domain)[8].

The Human Interconnection Model

Layer	Deployment
10. Legal / Regulatory	Agent Layers
9. Organizational Policy	
8. Individual Human	
7. Application	
6. Presentation	Services, Middleware, Operating System Layers
5. Session	
4. Transport	
3. Network	
2. Data link	
1. Physical Layer	Physical Layer

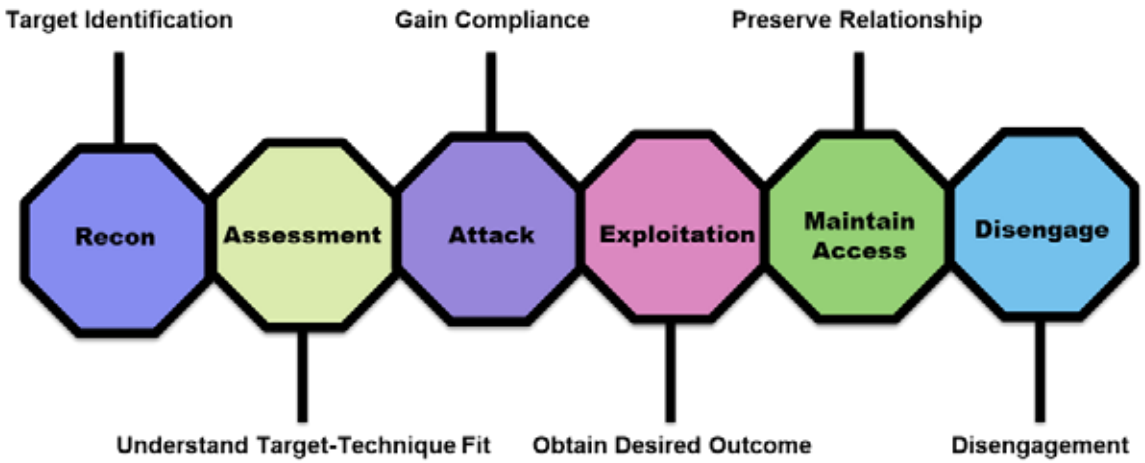
Some in the security research community have suggested the creation of a Human Interconnection Model (HIM)[9,10,11] to extend the OSI Model (layers 1-7)[12], to include an 8th layer for (uncoordinated) humans, a 9th layer for organizational policy, and a 10th layer to describe legal and regulatory systems. The distinguishing characteristics between these layers is the "protocol" by which they operate. Individual human beings and unorganized mobs operate differently than bureaucratic organizations with infrastructures and established policies. While corporations wield tremendous power both domestically and internationally, governments (nation states) currently have monopolies on the use of physical and kinetic force. These additional layers become increasingly significant as technology becomes more integrated with daily human interactions, leading to the emergence of *socio-technical systems* and when considering the security of *deterministic* versus *non-deterministic* systems. For example, many large language model (LLM) jailbreaking attacks (Layer 7) bear striking similarity to social engineering techniques applied against humans (Layer 8)[13]. Additionally, AI agents (Layer 7) pose a greater potential to present as an insider threat within an organization as their access and autonomy increase[14]. The emergence of smart contracts or similar techno-legal mechanisms are likely to interface with autonomous and semi-autonomous AI systems (Layer 10) leading to socio-technical legal frameworks and "smart" systems[15]. Finally, organizational policy changes (Layer 9) such as deciding to purchase certain supplies from vendors with chosen characteristics (lowest cost, most environmentally friendly) may be exploited by threat actors seeking to poison a supply chain.

When security considerations are embedded within an expanded framework (OSI Layers 1-7 and HIM Layers 8-10), an enhanced **Cognitive Security Risk Management Framework** emerges which offers tools and resources—as well as risk exposures and liabilitiesat each layer. For example, Layer 10 offers security professionals with legal tools to bring to bear against insider threats (Theft of Trade Secrets) but also presents potential exposures such as when ransomware groups threaten to report data breaches to regulatory agencies[16].

Layer	Tools/Resources	Exposures / Liabilities
Layer 10 (Legal, Reg.)	Economic Espionage Act, Theft of Trade Secrets (USC), FBI Investigation Capabilities	Legal requirements may also expose organization to liabilities. Example: SEC Breach Reporting, GDPR
Layer 9 (Org. Policy)	Organizational Policy & Procedures	Policy friction or rigidness may promote noncompliance Example: Shadow IT
Layer 8 (Ind. humans)	Training & Assessment	Insider threats (intentional, unintentional)
Layer 7 (Agentic AI)	Guardrails	Legal liability for AI agent actions
OSI Layers 2-6	Standard suite of cyber protection tools	Cybersecurity
Layer 1 (Physical)	Locks, seals, countersurveillance, training, OPSEC	Physical security

The Cognitive Attack Kill Chain

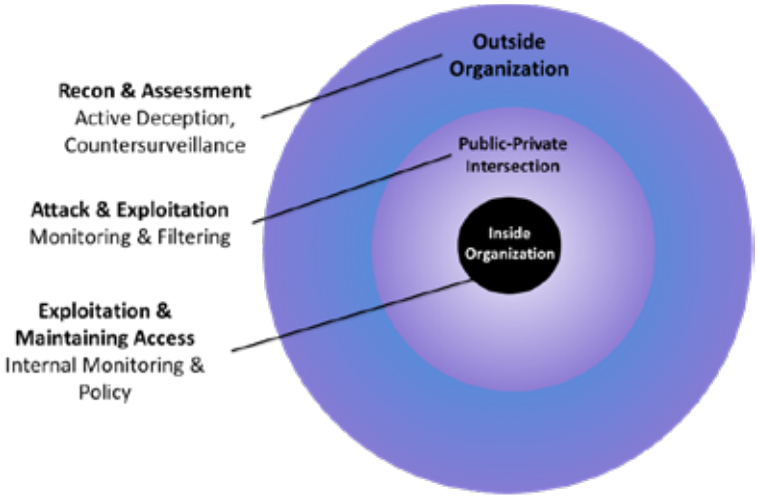
Threat actors seeking to manipulate targets with malicious intentions proceed through a series of steps not unlike the Cyber Kill Chain[17]. Like the Cyber Kill Chain, the cognitive analog suggests several points at which a cognitive attack may be disrupted. Initially, threat actors will engage in the *Reconnaissance* phase which seeks to identify a target suitable for exploitation. To meet this criterion a target must usually be accessible to the attacker, have access to a desired resource, and/or the ability to take the desired action. One reason that phishing attacks are so commonly used is that they offer criminals with potential access to anyone on Earth who has an active email account. If the phishing target does not have access to a bank account, or sufficient funds to interest a criminal, then that target will be ignored in favor of more desirable targets, even though the criminal has the ability to access them through email. During the *Assessment* phase, the next step in the kill chain, the attacker's goal is to uncover *target-attack fit*, meaning that they need to understand which technique will be effective against a particular target. This often manifests as a "pretext", the story that the attacker is presenting to the target. This may sometimes occur simultaneously with launching the attack itself, such as a 419 (aka advance fee, or Nigerian Prince) scam, in which targets are "filtered" according by their willingness to respond. In the *Attack* phase the attacker's goal is to gain the target's compliance by using the method(s) identified in the previous phase. If the *Exploitation* phase is successful, the attacker will have gained the target's compliance through the attack, and the attacker will now need to decide whether to terminate the engagement or continue. If the attacker decides to continue *Maintaining Access* in the next phase of the attack, their goal will be to preserve and solidify their relationship with the target. This phase is most frequently observed in long-term engagements such as cat fishing, romance scams, or espionage agents being operated by intelligence officers.



The final phase of *Disengagement* is most critical when the attacker may wish to reengage with the target later or wishes to remain undetected. This phase will be critical for an industrial spy who elicits sensitive information from a target in a non-alerting manner.

Phase	Objective	Examples of Actions
Recon	Target Identification	OSINT, Background Research
Assessment	Understand Attack-Target Fit	Surveillance, In-Depth Research
Attack	Gain Compliance	Social Engineering TTPs, Influence Messaging
Exploitation	Obtain Desired Outcome	SE TTPs focused on compliance
Maintain Access	Preserve Relationship	Commitment & Consistency techniques
Disengage	Avoid Detection (usually)	Use of cover story for exit.

Defending The Enterprise Against Cognitive Attacks



The previously described **Cognitive Attack Kill Chain** illuminates risk exposures that attackers experience at each stage of an attack. Enterprise defenders who understand these exposures may leverage these to thwart cognitive attackers at different stages using appropriate techniques for each stage.

"Early" stages of the Cognitive Attack Kill Chain, such as Reconnaissance and Assessment might be considered as "far" outside the organization's boundary because these activities may be conducted with a very low detection signature. It is nearly impossible for defenders to prevent these activities from occurring, which is why deception and countersurveillance are the most effective methods to counter-attacks during these stages. Employee training to counter these phases should focus on identifying elicitation and surveillance techniques.

The "middle" stages of the kill chain involve launching the Attack and Exploiting the target. Attackers operating at these stages are most vulnerable to traditional countermeasures such as email filtering and authentication methods. Employee training to counter these stages follow typical security awareness templates of identifying social engineering techniques.

The "late" stages of cognitive attacks involve further Exploiting and Maintaining Access to the target. These stages will typically execute within the perimeter of the organization's control and therefore internal monitoring needs to play a major role in detection and mitigation. Employee training to detect insider threats often yields mixed results since most people do not have the desire to "turn in" their peers or coworkers. Training to detect insider threats should be oriented toward management and human resources professionals. An "open-door" or "safe space" policy for reporting is paramount for countering these threats. It is noteworthy that several public reports have surfaced of ransomware gangs attempting to recruit employees to deploy malware within their organizations, several of these were thwarted because the targeted employees reported the recruitment attempt(s)[18]; however,

several reports have also emerged which suggest that a combination of incentives (money) and coercive (extortion) are currently being used. It is very difficult for an employee to report that the recruitment attempt occurred while they were “interviewing” for another job.

Checkers versus Chess: Learning to think Strategically in Cognitive Security

A threat actor always has a reason or an objective they are trying to achieve when launching an attack. Most frequently this motivation is money, but recent estimates suggest that the objective of one out of twenty attacks is motivated by espionage[19]. Espionage as the primary motivation substantially increases within industries which have access to higher value information, such as national defense information; however, this trend may be changing as nation-states appear to be increasing their collaboration with cybercriminal groups[20, 21].

Level	Activity	Indicators of Engagement
Tactical	Individual interactions	Social engineering TTPs
Operational	Coordinated actions	Surveillance indicators
Strategic	“The Grand Plan”	Indicators of engagement at multiple layers (1-10) over an extended period of time.

Scenarios

The following sections provide additional context and sources for the RSAC 2025 Learning Lab tabletop exercise scenarios. Whenever possible, these scenarios attempted to draw from real-world occurrences.

Scenario 1: X Marks the Sport

Politically motivated activism has been increasingly shifting focus toward private corporations over the past 20 years and there is currently no indication this trend to shift in the foreseeable future. A key point in this exercise is that that activist/hacktivists appear to be escalating the seriousness of their actions, and the trajectory is trending from cyber defacement and minor mischief toward physical damage and potentially threats to personal safety. This trajectory argues against allowing executive staff to appear in person to pay a ransom as suggested in the tabletop exercise.

The murder of United Healthcare CEO Brian Thompson appears to have required a significant degree of pre-planning, online reconnaissance, and open-source intelligence (OSINT) to ascertain patterns of life. It is likely that future attacks will follow similar methods and may adopt a more active posture by attempting to induce the target to be physically present at a specific time and location. Periods of increasing disparity in wealth distribution has historically resulted in higher rates of violence directed toward high-net worth individuals and high-level executives. The recent historical trend toward greater wealth disparity suggests additional attacks targeting high-net worth individuals are likely to occur in the future, which is supported by the significant amount of public support for Luigi Mangione[22] along with similar attacks which have recently occurred[23,24,25,26,27,28] . It is likely that some portion of these attacker will move beyond mere surveillance and will begin inducing targets to move to desired

locations in an attempt to “put them on the X” at a time and location advantageous to the attacker. Such actions were described previously in the Multi-Domain Security section above.

This tabletop scenario was intended to demonstrate the interrelationship between the cyber, physical, and cognitive security domains and the necessity to consider the implications each on the others.

Scenario 2: Ghost in the System

This scenario was intended to bring attention to new evolutions in insider threats, with the potential for AI agents to go rogue within an organization in response to a signal(s)[14], the use of proxies to obfuscate the true insider threat, and the use of readily available[29] capabilities to impersonate real people to become fake employees[30]. Deepfake and other forms of synthetic media are rapidly advancing threat actor capabilities to impersonate known individuals with very small voice samples or individual photos.

Scenario 3: Multi-Level Marketing

This scenario was based on a real series of events involving corporate espionage between the Nestle Chocolate and Mars corporations and was documented in the book Broker, Trader, Lawyer, Spy: The Secret World of Corporate Espionage by Eamon Javers[31]. The objective of this scenario was to get defenders thinking at the strategic and operational levels rather than becoming fixated on tactical level cognitive attacks.

Sources

1 Finance worker pays out $25 million after video call with deepfake ‘chief financial officer’

https://www.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html

https://archive.is/1KVVQ

2 The Cybersecurity 202: Iran indictments show even U.S. intelligence officials are vulnerable to basic hacking schemes

https://www.washingtonpost.com/news/powerpost/paloma/the-cybersecurity-202/2019/02/14/the-cybersecurity-202-iran-indictments-show-even-u-s-intelligence-officials-are-vulnerable-to-basic-hacking-schemes/5c6462af1b326b71858c6b78/?utm_term=.ff7574e12ceb

https://archive.is/zHSro

3 Robbers Use Pokemon Go to Lure Victims

https://www.facebook.com/permalink.php?story_fbid=1272368709470229&id=180316078675503#

https://archive.is/ANe4M

4 Armed muggers use Pokémon Go to find victims

https://arstechnica.com/gaming/2016/07/armed-muggers-use-pokemon-go-to-find-victims/

https://archive.is/bEiVK

5 Deputies warn of dangers from child predators playing Pokemon Go

https://www.wxyz.com/news/deputies-warn-of-dangers-from-child-predators-playing-pokemon-go

https://archive.is/8YdUo

6 An artist wheeled 99 smartphones around in a wagon to create fake traffic jams on Google Maps

https://www.businessinsider.com/google-maps-traffic-jam-99-smartphones-wagon-2020-2?op=1

https://archive.is/1L6Bn

7 Waze sent commuters toward California wildfires, drivers say

https://www.usatoday.com/story/tech/news/2017/12/07/california-fires-navigation-apps-like-waze-sent-commuters-into-flames-drivers/930904001/

https://archive.is/LvuRS

8 Raman, G., AlShebli, B., Waniek, M., Rahwan, T., & Peng, J. C. H. (2020). How weaponizing disinformation can bring down a city’s power grid. PloS one, 15(8), e0236517.

https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0236517&type=printable

9 Engineering Security Solutions at Layer 8 and Above

https://web.archive.org/web/20120709142600/http://blogs.rsa.com:80/curry/engineering-security-solutions-at-layer-8-and-above/

10 Teaching Cybersecurity Policy
https://www.schneier.com/blog/archives/2018/12/teaching_cybers.html

11 Schneier: Government, Big Data Pose Bigger 'Net Threat than Criminals
https://www.schneier.com/news/archives/2012/02/schneier_government.html

12 https://en.wikipedia.org/wiki/OSI_model

13 Zeng, Y., Lin, H., Zhang, J., Yang, D., Jia, R., & Shi, W. (2024). How Johnny Can Persuade LLMs to Jailbreak Them: Rethinking Persuasion to Challenge AI Safety by Humanizing LLMs. arXiv preprint arXiv:2401.06373.
https://arxiv.org/pdf/2401.06373

14 Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., & Perez, E. (2024). Sleeper agents: Training deceptive llms that persist through safety training. arXiv preprint arXiv:2401.05566.
https://arxiv.org/pdf/2401.05566

15 We Must Bridge the Gap Between Technology and Policymaking. Our Future Depends on It
https://www.schneier.com/essays/archives/2019/11/we_must_bridge_the_g.html

16 Ransomware gang files SEC complaint over victim's undisclosed breach
https://www.bleepingcomputer.com/news/security/ransomware-gang-files-sec-complaint-over-victims-undisclosed-breach/
https://archive.is/op9fi

17 Cyber Kill Chain
https://www.lockheedmartin.com/en-us/capabilities/cyber/cyber-kill-chain.html

18 Russian to be deported after failed Tesla ransomware plot
https://apnews.com/article/europe-russia-technology-business-government-and-politics-4860aa6008eaba24b604f9aa6fb9b138
https://archive.ph/xfrHh

19 2025 Data Breach Investigations Report
https://www.verizon.com/business/resources/T41e/reports/2025-dbir-data-breach-investigations-report.pdf

20 Microsoft report highlights nation-state actors' collaborations in cyberattacks
https://backendnews.net/microsoft-report-highlights-nation-state-actors-collaborations-in-cyberattacks/

21 Microsoft Digital Defense Report 2024
https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/final/en-us/microsoft-brand/documents/Microsoft%20Digital%20Defense%20Report%202024%20%281%29.pdf

22 Luigi Mangione Has Become A Social Media Folk Hero
https://www.forbes.com/sites/petersuciu/2024/12/12/luigi-mangione-has-become-a-social-media-folk-hero/

23 https://www.linkedin.com/pulse/why-executives-new-prime-targets-cyberattacksand-how-fight-demirsoy-gnksc/

24 https://www.thetimes.com/uk/crime/article/brazen-thief-snatched-jewellery-worth-more-than-10m-qfr8whr05?utm_source=chatgpt.com&region=global

25 https://www.thetimes.co.uk/article/boohoo-boss-quit-amid-corporate-espionage-and-stalking-of-executives-jvzrkj7x6?utm_source=chatgpt.com

26 https://people.com/unitedhealth-ceo-brian-thompson-targeted-gunman-lying-wait-8755102

27 https://nypost.com/2024/10/24/world-news/australian-billiona/

28 https://www.kptv.com/2025/02/25/ceos-lake-oswego-home-possibly-targeted-shooting/

29 The FAIK Files Decoding AI Deception in Our Hyperreal World with Perry Carpenter | CSI #41
https://www.youtube.com/watch?v=7X82u2SpvIc

30 KnowBe4 Issues Warning to Organizations After Hiring Fake North Korean Employee
https://www.knowbe4.com/press/knowbe4-issues-warning-to-organizations-after-hiring-fake-north-korean-employee

31 Javers, E. (2010). Broker, trader, lawyer, spy: The secret world of corporate espionage. Harper Collins.
https://www.harpercollins.com/products/broker-trader-lawyer-spy-eamon-javers?variant=32207471673378