The Agentic AI Army That Never Was: Projecting LLM Swarm Narratives with 'Noisy' LLM Sock Puppets and Whaley's *Theory of Outs*

Tim Pappa[1] and Christopher Williams[2]

[1]Independent researcher
[2]Independent researcher
timothypappa@gmail.com

Abstract

This short position paper suggests there may be greater deception and influence of an attacker's perceptions of a fictional 'Agentic AI Army' swarm of LLM sock puppet network defenders than deploying real LLM agent swarms. We model a counterintuitive industry approach integrating Whaley's lesser-known Theory of Outs and "turnabout" deception techniques to encourage a human or LLM attacker's discovery of deception on an industry network. While we recognize that the knowledge of real or imagined deception can deter an attacker, we also recognize that attackers may demonstrate greater confidence on a network after discovering what appears to be deception artifacts. We visualize how 'noisy' LLM sock puppets inside of a network that prompt optimized query returns on their content and placement on the network could draw attackers to later stage deception functions and effects and enhanced defender alerting and analysis on human or LLM attacker interaction with those deception functions. We find in anecdotal operational research that highlighting 'noisy' sock puppet content enhances high-fidelity detection. We frame these findings using this integrated industry model in the context of LLM swarm narratives for deception. There has been an increasing concentration on swarming as a military technique and military strategy, as modern military conflicts continue to adapt to irregular warfare environments. The renewed concentration on developing and integrating swarm intelligence with LLM agents continues to face limitations, in terms of simulating natural swarm behaviors and operating autonomously as part of a decentralized model. This short position paper proposes a more immediate deception and influence effect, namely projecting fictional LLM swarm narratives suggesting there is an 'Agentic AI Army' assisting human defenders. We use organizational perception management as a design framework to visualize a deception and influence narrative communicating this fictional narrative using 'noisy' LLM sock puppets and our integrated model of Whaley's Theory of Outs and "turnabout" deception techniques.

Introduction

The proposed integrated modeling in this short position paper is foundationally based on an arguably counterintuitive operational approach to industry cyber deception, where deception is intentionally 'noisy' and staged to be discovered. Our recent experience in industry cyber deception operations has involved more of an organizational approach to deception functions and effects, where generally there is less direct engagement with attackers than we previously experienced when supporting law enforcement operations online in prior roles. Because there is generally less direct engagement with attackers, there is often less observed attacker behaviors. Designing deception functions and effects without a baseline of behavior is arguably difficult. There must be some application of behavioral and cognitive frameworks in the absence of demonstrated attacker behaviors.

This short paper is also foundationally about influencing human attackers' real and imagined perceptions of an organization's LLM agent defender swarm capabilities with deceptive fictional content and communication. Organizational perception management explains how people inside and outside organizations form real and imagined perceptions of that organization. The projection of fictional LLM swarm defender narratives suggested in this paper is in response to considerable public commentary regarding LLM capabilities, with disagreements among cybersecurity industry practitioners and the broader public about what LLMs can do and what LLMs cannot do. This kind of confusion about what is real creates opportunities for deception and influence.

We define deception as an intentional distortion or manipulation of the way an attacker thinks and feels or processes information. We conceptualize LLM agents in the context of this paper as *decoying* deception techniques. Deception frameworks generally characterize *decoying* as a form of *simulation* (Whaley and Bell, 1982, 2016). We define *decoying* as showing the false by masquerading, but we have established in development of our cyber deception operations that *decoying* is different than *inventing*, which is another form of *simulation*. We consider inventing much

more static, whereas *decoying* is more dynamic or more responsive to interaction with an attacker. While the scope of this extended abstract does not include an extended discussion of foundational deception principles, we do want to note that deception *simulation* and *dissimulation* must be designed and understood as simultaneous (Whaley, 1974). This design approach complements any conceptualization or imagination of the swarm behaviors of an imagined or fictional swarm of LLM agent defenders, who are simultaneously *simulating* that an organization has an LLM swarm capability but also *dissimulating* that an organization has no LLM swarm. We want to generally characterize our experience and approach to developing and operationalizing LLM infrastructure. We have demonstrated some progression in refining an LLM to generate content for our operations. We also work for an industry organization that has been intensively promoting integration of artificial intelligence functions and LLM infrastructure, such as LLM 'Super Agents' that orchestrate systematic organization functions. We understand the difference between base LLMs designed to generate responses to prompts and the kind of artificial intelligence (AI) we characterize in this short paper – which are autonomous LLM agents powered by LLM base orchestrators, and commercial AI platforms we integrate to advance some of our development and operations. This paper builds on "adaptive cyber deception" research that encourages design based on observed behaviors, but we are integrating the design of imagined LLM agents to influence the perceptions of human attackers (Aggarwal et al., 2024; Gonzalez et al., 2020; Cranford et al., 2021). We have integrated this cyber deception design approach with other historical military techniques, such as ambushes and raids (Pappa, Dirie, and Bradford, 2024).

The examples of swarming behaviors are plentiful, not just in historical and contemporary military conflicts but in the study of instinctual collective animal behaviors. Mongol fighters have been highlighted often in military work on swarming because of their effectiveness in military campaigns, using swarming not only as a tactical move in battle but as a foundational part of their overall strategy. Mongol leadership enabled a decentralized command system among subordinate commanders, so they had more freedom to make decisions during a battle with their fighters. The Mongol army consisted almost entirely of cavalry, too, which facilitated greater mobility and situational awareness for Mongol leadership than most enemies. The information collected in swarming techniques was key to relaying messages and battlefield updates promptly. German submarine warfare has also been referenced often in more contemporary military work on swarming. German U-boat "Wolfpacks" would swarm Allied troop transport ships and commercial cargo ships trying to cross the Atlantic in the early years of World War II before there was Allied air cover support for the entire crossing. German U-boat packs of five or more submarines would swarm these convoys that included Allied destroyers and confuse and distract those destroyers until they isolated less protected or unarmed shipping to destroy those ships. The goal of swarming in many cases is less the physical destruction of an enemy and more the "disruption of its cohesion". Many Allied ships sunk, and many lives were lost, but the psychological effect may have been greater in the public's morale and impressions of the Allied support effort and the Allied struggle to overcome these U-boats. This paper reflects a reinvigorated study of military swarming techniques in the past two decades as military conflicts continue to occur in more irregular warfare contexts and less conventional environments, but we will be exploring what the recent studies of LLM swarm capabilities are and if a deception and influence approach with fictional narratives about LLM swarm defenders could influence human and LLM attackers more than deploying real operational LLM agents.

This short paper is organized as follows. First, this paper will present a working characterization of an integrated "turnabout" and "Theory of Outs" model of cyber deception design. Second, this paper will briefly discuss related work on recent research into agentic LLM swarms and the challenges that development continues to face. Third, this paper will feature two visualizations. The first visualization will describe our proposed integrated model in a scenario. The second visualization will apply the same integrated model but within a design framework of organizational perception management to describe the communication of a fictional 'Agentic AI Army' narrative.

A Working Characterization of Integrated "Turnabout" and "Outs" Deception Planning

Whaley (2016) characterized "turnabout" as a context where deception appears to have been discovered, and there is an opportunity to respond with deception. "Turnabout" was a historical and practiced technique in military deception, but Whaley found that planning for "turnabout" in contemporary deception operations in the military appeared to be considerably rare in contrast to the practice of "turnabout" in historical military examples.

Whaley wrote about what has been called a "Let's Double Back trick" or a "double back" as another representation of "turnabout". This generally means someone has 'doubled back' to a starting point in an ongoing deception, whether that deception was planned or not. The audience or target of that deception may presume that someone in some pattern is following a sequence of activity that will lead to an anticipated outcome, rather than returning to a starting point. The "double back" is surprising, then. But deception planners planned for it. Whaley referred to a deception operation in the 1970s where the Central Intelligence Agency had been

collecting imagery of the Soviet Union from their KH-11 "Keyhole" satellite, searching for potential SALT agreement violations related to missile production. Because the Soviets were aware of this collection, they would often relocate or hide missile infrastructure. When the CIA later incorporated an unknown new design feature in the KH-11 satellite, where collection was transferred to another satellite and then transmitted to the ground rather than directly from the KH-11 to ground, the Soviets could no longer detect any data imagery transmissions from the KH-11 to ground. The Soviets' conclusion appeared to be that the satellite was "dead". The Soviets stopped hiding their missiles during these flyovers because they believed the satellite was not collecting imagery. The CIA, however, continued collection and transmission, until the Soviets later discovered from an asset that the CIA had modified the satellite.

Whaley referred to another example of "double back" from an operation known as MOONSHINE from World War II. The British had begun distorting German radar detection by using older models of radar emulation technology. When the British began 'doubling back' or reintroducing older versions of their radar technology, the Germans struggled to accurately detect enemy aircraft. The number or scale of approaching British aircraft was distorted because the Germans had configured their detection systems to look for more advanced British radar hardware. The Germans eventually discovered this "double back" and adjusted their detection efforts. About a year after that discovery, the British reintroduced this early version of their radar emulation of an older model again, which again distorted German detection efforts because they expected to see more advanced versions.

Whaley differentiated his *Theory of Outs* as planning for "alternative goals" for a deception if that deception is discovered or if the deception appears to have failed. Whaley wrote that deception rarely fails because there is generally some effect because of deception where the target is influenced in some manner, but deception planners in his experience consistently did not include options in their deception operations for accomplishing the deception or influence goal in some other way if the deception was discovered. Whaley wrote about the "Delayed Message Trick" where the true operational plans are transmitted on a channel the target is monitoring, but the transmission is slowed or delayed in a manner where the target will be unable to react effectively. Whaley wrote that this approach is rare and should be used sparingly, but this is an example of planning for an "out" by managing the transmission of the operational plans but perhaps keeping those plans vague and making it appear there are technical issues with the target's monitoring infrastructure. There is still flexibility in this deception design to hold the target's attention and influence them, but to control the transmission or communication of that deception content.

This short position paper characterizes a working approach to deception planning or deception design integrating "turnabout" and this theory of "Outs", where planners include the discovery of the deception into their design. Whaley may have intended these approaches or techniques in deception to be separate given the context or situation, but we suggest that in a naturalistic offline and online network attack environment there should be an integration of these techniques and layered deception design.

Developed Enough to Believe We Have Natural LLM Agent Swarms

Military swarming has been characterized as seemingly amorphous, but this approach is deliberate in structure and coordination (Arquilla and Ronfeldt, 2013). This approach applies to development of operational LLM agents, too.
The research exploring swarm intelligence in the past year has concentrated mostly on the capability of LLM agents to operate in a decentralized structure or system that resembles natural swarm behaviors and swarm intelligence. Ruan et al. (2025) introduced the constraints found in natural swarm environments, such as limited communication and "strict decentralization", to see how LLM agents could effectively coordinate and communicate their behaviors. Their findings suggested that the LLM agents modeling in their study did demonstrate some coordination, although these LLM agents struggled with collective behavior under those same strict natural swarm environment factors or norms. Randevik and Petersson (2025) proposed the replacement of an LLM agent with several agents acting as assistants to the LLM agent, demonstrating swarm behavior while functioning based on its own set of instructions and goals. When comparing the prompts from a human user and ChatGPT, the AI agent assistants performed better with more detailed instructions from the human user. The researchers argued that this need for further prompting clarity suggests that AI swarms may not be as viable for general use as perhaps anticipated. Jimenez-Romero, Yegenoglu and Blum (2025) examined two approaches to operationalizing LLM agents in a multi-agent environment, both focusing on the role of prompt engineering in guiding LLM agent behaviors. The first approach simulated an ant colony where LLM agents representing individual ants were prompted with defined conditions and tasks, such as retrieving food. The researchers wrote that this method "allows for precise control over agent actions, enabling a rule-based system where each agent's behavior is explicitly dictated by the LLM-generated instructions". The second approach used less structured, principle-based prompts simulating a flock of birds. The prompts in this second approach relied on the LLM's understanding of complex concepts such as "flocking dynamics

and self-organization". This approach enabled the LLM agents to manage these behavioral patterns.

Rahman, Schranz, and Hayat (2025) also explored if an LLM swarm framework where agents coordinate through natural language prompts can demonstrate similar constraints and qualities of natural or classic swarm intelligence, like decentralization and scalability. Their findings suggested that while these LLM-powered swarms can demonstrate swarm behaviors, there is considerable infrastructure resources and support required. For example, one of the researcher's simulations required approximately 300X more computation time than standard computing data.

This research also suggested that LLM-driven swarms are not yet prepared to operate in real-time systems. Certainly, this research can change within months and there is a growing body of LLM swarm research challenging these same findings, suggesting for example that multi-LLM collaboration can train native swarm models better than prompt refining and demonstrate the natural swarm behaviors better than prior models (Feng et al., 2025). We argue that there is limited operational value at this stage in industry development to attempt to engineer an agentic LLM swarm when giving the impression of an agentic LLM swarm even if that swarm is fictional can still be effective.

Visualizing a 'Noisy' LLM Sock Puppet Deception Approach Integrating "Turnabout" and "Outs" Modeling

In the first visualized scenario, we are operating an LLM agent sock puppet internally on a company enterprise. This LLM sock puppet is presumably assisting software developers working on a new enterprise platform that another sock puppet we are operating that appears to be a junior DevOps engineer has been blogging about on an external social media platform. The engineer sock puppet has written generally about his recent projects, but there have been some unique keywords he has included in his blog content that refer to a unique project name and references to this LLM agent that is presumably assisting this engineer and other software developers. The engineer blogger might also suggest that much of the company data on this new platform is proprietary and the functions of the platform would be considered sensitive.

Even if a human or LLM agent attacker in this visualized scenario obtained unauthorized access to this company's network simply by obtaining basic user account credentials, there will likely be some effort inside the network to verify whether the engineer sock puppet account blogging externally is a company employee and if the unique keywords and references to these fictional projects and the sock puppet LLM agent guide the attacker to our deception environment. When the attackers in this scenario begin to query certain terms or keywords associated with this sensitive platform and the LLM agent, the content and landing page return should have been optimized to ensure the attacker continues to our deception environment associated with those queries and content. Because there has been consistent content sharing by this sock puppet, there may have been prior false positives from other company employees curious about these projects or employees questioning the content he or she is posting. This is what makes this sock puppet content 'noisy', but this approach is also a method to lure human or LLM agent attackers who are refining their search based on our manipulated query and search return design. Network defenders in this scenario may have also placed simple decoys along this possible attack pathway for an attacker, so that the attacker will avoid what appear to be oddly placed honey files with likely alerting and perhaps obvious naming conventions on the files, such as "sensitive platform data_confidential_Q1". We recognize that this kind of naming convention may still appeal to attackers who will try to access this kind of honey file because of the naming convention or description of the data in the file in the title, however, this kind of approach would still demonstrate our proposed integrated modeling of "turnabout" and "Outs" deception planning. Network defenders want the attacker in this scenario to believe he or she has discovered these honey files in a plausible attack pathway or that an LLM agent attacker has logically exhausted protocol or sequence in artifact discovery and examination, so that these imagined attackers will either trip those decoys and return with other credentials in the future or they will avoid these honey files and continue their movement, perhaps confident they have detected this deception or that they have completed prompted instructions left for the LLM agent attackers. At any point in this sequence, the attacker could discover this deception environment and abandon the network. We would argue the attacker would likely avoid future attempts.

In this scenario, once a human attacker finds content of interest from his or her queries of keywords or terms of interest and the search return includes the engineer sock puppet's content or the references to the LLM agent sock puppet, the human attacker may scrutinize this content less because he or she is familiar with this deception storyline, but unaware this is deceptive. The indexed content may include a hyperlink to another site on the network that suggests access to the data the attacker is seeking. When the attacker clicks on this link, he or she is transferred to a site where he or she can enter his or her stolen user credentials. That access attempt will be successful, but there will be alerts. Then the attacker will be transferred to another controlled portal with further warning on that site regarding confidential access. When the attacker attempts to sign into this portal with the same credentials, the SOC will be alerted immediately. This incremental sequence of deception effects and function limits the number of false positives, discouraging employees

from attempting to explore and access these sites and content. This sequence of deception functions also allows the SOC to observe a user's intent and to capture a high-fidelity alert. These steps could be modified for LLM agent attackers as well.

Visualizing a Fictional Swarm Narrative with an Organizational Perception Management Design Framework

In the second visualized scenario, we apply the design of a 'noisy' LLM agent sock puppet within the framework of organizational perception management. Organizational perception management has been characterized as "phenomena" because of how dynamic impression formation is. Elsbach (2003) defined the phenomena of organizational management as "symbolic actions" by people within an organization to influence the perceptions of that organization inside and outside. Elsbach preferred the use of "perception management" rather than "impression management" because there are broader considerations for how people represent an organization and communicate with people inside and outside of an organization. Elsbach found that many organizations simultaneously demonstrated "symbolic actions" that reflected images or reputations of that organization that influenced how someone formed real and imagined perceptions of that organization.

In this scenario, we work for an industry organization that has been intensively demonstrating integration of artificial intelligence functions and LLM infrastructure, so the public suggestion of 'we have LLM Agent swarm defenders' is plausible as a narrative. In this design visualization, we would identify public content and commentary for sock puppets to comment on or write about. This public content and commentary provide the authentic storyline or narratives for us to add plausible variations to that storyline or narrative that are difficult for people including attackers outside of the organization to verify. In this scenario, we control a sock puppet account on GitHub that we have periodically posted some modified work that appears original and we have forked existing shared work from researchers focusing on LLM refinement and LLM agent development. This provides some plausible context for a storyline in which an industry engineer maintains a personal GitHub account that reflects his or her personal and professional interests, even if he or she is prohibited from posting company data or projects on his public GitHub. Given the industry reporting on human and LLM agent attackers scanning and scraping public GitHub repositories looking for vulnerabilities and people of interest to target for credentials, attackers may find this deception sock puppet account with unique keyword searches related to LLM agentic swarm research and development. This sock puppet storyline would reflect a real and imagined narrative of this industry's LLM and artificial intelligence concentration, which might make the storyline content appear more authentic or plausible and may encourage an attacker to explore this narrative further. The GitHub sock puppet account could include naming conventions that are unique to the 'noisy' sock puppet LLM or collection of LLM agentic swarms inside the network that a human or LLM agent attacker may search for if that attacker is able to enter the network with a compromised user account. The norm of these kinds of artifacts or narratives related to LLMs within the network may limit false positives and the specific searching or query keyword criteria would also limit false positives from common network users, but this 'noisy' fictional LLM sock puppet swarm and narrative content would still likely misdirect attackers to us.

Discussion

These visualized scenarios only briefly discuss the operationalization of deception in the form of a fictional narrative that could influence the decisions a human or LLM agent attacker might make, in terms of how the attacker might attempt to break into our organization's network and what they might search for if they access our network with stolen credentials, for example. There could be unique phrases or references to uniquely named programs we include in the sock puppet's public content and communication that would enable us to determine an alert on that same queried phrase or reference on some of our internal platforms was likely from an attacker, not a valid user. This approach extends the influence of a deception environment outside of an internal network to open web environments and platforms where human and LLM agent attackers routinely conduct reconnaissance. Misdirection starts outside of the network instead of inside the network where the attacker already has gained access. Generally, we try to *pre-suade* attackers outside the network and then persuade or influence them inside the network with our 'noisy' sock puppet content (Cialdini, 2020). We generally believe we can compound the influence of our deceptive narratives if our deception is discovered; much of the research in the past half decade on deception online has found that even if there is no deception that informing an attacker there is deception or there might be deception is sufficient to influence their behaviors online (Ferguson-Walter, 2024; Ferguson-Walter et al., 2021). If an attacker discovers that their pathway of reconnaissance and unauthorized access was based on a fictional narrative, then we believe a human attacker or an LLM agent attacker might logically determine that other functions or effects the attacker has executed or observed may have also been manipulated in some way. This does not include the influence of affect or emotions for a human attacker when they dis-

covered deception, or for an LLM agent attacker when it determines there has been some attempt to misdirect it and it becomes more cautious. We would argue that not only would an attacker attempt to exit the network, but the introduction of uncertainty and ambiguity in this network's methods of defense and deception may deter future attempts by the same or similar attackers. The attackers operating the LLM agent attackers might make the same conclusion. This kind of attacker interaction with fictional narratives and deception functions and effects embedded in storyline content could markedly influence the attacker's real and imagined perceptions of the capabilities and vulnerabilities of an organization's enterprise network defense.

References

Arquilla, J. and Ronfeldt, D., 2013. Swarming and the Future of Conflict.

Cialdini, R., 2020. *Pre-suasion*. ReadinGraphics.

Edwards, S.J., 2000. Swarming on the Battlefield: past, Present, and Future.

Feng, S., Wang, Z., Wang, Y., Ebrahimi, S., Palangi, H., Miculicich, L., Kulshrestha, A., Rauschmayr, N., Choi, Y., Tsvetkov, Y. and Lee, C.Y., 2024. Model swarms: Collaborative search to adapt llm experts via swarm intelligence. *arXiv preprint arXiv:2410.11163*.

Ferguson-Walter, K.J., 2024. An empirical assessment of the effectiveness of deception for cyber defense.

Ferguson-Walter, K.J., Major, M.M., Johnson, C.K. and Muhleman, D.H., 2021. Examining the efficacy of decoy-based and psychological cyber deception. In 30th USENIX security symposium (USENIX Security 21) (pp. 1127-1144).

Jimenez-Romero, C., Yegenoglu, A. and Blum, C., 2025. Multi-agent systems powered by large language models: applications in swarm intelligence. *Frontiers in Artificial Intelligence*, *8*, p.1593017.

Pappa, T., Dirie, A. and Bradford, J., 2024, October. Applying Models of Historical Mujahideen Ambushes and Raids to Cyber Deception Practitioner Design. In *MILCOM 2024-2024 IEEE Military Communications Conference (MILCOM)* (pp. 1-6). IEEE.

Randevik, N. and Petersson, D., 2025. Evaluation of an AI Assistant Swarm of LLM-based Agents. *LU-CS/HBG-EX*.

Rahman, M.A.U., Schranz, M. and Hayat, S., 2025. LLM-Powered Swarms: A New Frontier or a Conceptual Stretch?. *arXiv preprint arXiv:2506.14496*.

Ruan, K., Huang, M., Wen, J.R. and Sun, H., 2025. Benchmarking LLMs' Swarm intelligence. *arXiv preprint arXiv:2505.04364*.

Whaley, B., 1980. A Typology of Misperception or The Ways We Can Be Wrong. *Unpublished manuscript draft*.

Whaley, B., 1982. Toward a general theory of deception. *The Journal of Strategic Studies*, *5*(1), pp.178-192.

Whaley, B., 1974. Deception: Its Decline and Revival in International Conflict. *Unpublished manuscript draft*.

Whaley, B., 2016. *Turnabout and Deception: Crafting the Double-cross and the Theory of Outs*. Naval Institute Press.

Yuill, J., Denning, D. and Feer, F., 2007, January. Psychological vulnerabilities to deception, for use in computer security. In *DoD Cyber Crime Conference* (Vol. 2007).