HoneyContent: Wrapping Deception Storyline Content to Warrant Human and LLM Agent Attacker Evaluations of Deception Functions and Effects

Tim Pappa[1] and Darin Roberts[2]

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

Abstract

This position paper suggests that presenting deception storyline content could *warrant* or influence human and Large Language Model (LLM) Agent attackers' evaluations of deception functions and effects as more authentic. While there is growing attention to the need for cyber deception design that is adaptive to the cognitive and behavioral vulnerabilities of attackers, there has been limited discussion of how content and communication presenting deception storylines could influence how attackers evaluate the authenticity of real and imagined deception functions and effects on a network or host. There is perhaps even less understanding of how deception storyline content could influence LLM Agent attackers. We introduce *HoneyContent*, our practiced industry cyber deception model for designing and creating deception storyline content to 'wrap' or embed our deception functions. We will introduce *warranting theory* and *signaling theory* as the foundational content and communication frameworks we integrate to design and create our deception storyline content. We will also discuss recent findings from researchers modeling LLM Agent behaviors in response to introduction of content to misdirect their functions, namely related to informing these LLM Agents that a task was completed or that there is no need to perform a task. We recognize the similarity to presenting deception storyline content to influence humans and LLM Agent attackers. We will visualize the use of *HoneyContent* based on an industry cyber deception scenario that demonstrated anecdotally how this integrated design and content creation model could make deception functions look more real. We suggest that the use of generative artificial intelligence and LLM Agent defenders could further enhance warranted evaluations of these deception storylines and the embedded or wrapped deception functions and effects, because this presentation of deception storyline content is based on the behavioral vulnerabilities of these attackers.

Introduction

This position paper is foundationally based on how attackers evaluate and respond to content online. As industry cyber deception practitioners, we find that much of the discussion and research on cyber deception and cybersecurity has focused on how attackers try to exploit network infrastructure or how attackers respond to attempts with software tools and platform technologies to deter or disrupt them. What we rarely come across are demonstrations of the use of content to enhance the authenticity of deception.

We introduce two frameworks of content and communication evaluation to explore how the presentation of deception storylines can influence human and LLM Agent attacker evaluations of deception content and functions.

Warranting theory is a framework of content evaluation where people make determinations of whether the content they are viewing has been manipulated by a content creator. Signaling theory shares some similarities to warranting, but this framework is more concentrated on how people make determinations of how reliable someone is, based on what they are doing and how they are communicating. This paper will discuss how these frameworks can be integrated to model a content design and creation process that we have increasingly practiced in our development of industry cyber deception operations. We recently started modeling this design to influence LLM Agent attackers.

Because the scope of this paper is concentrated on deception storylines including content and communication, this paper will only briefly characterize the foundations of deception and how deception can influence evaluations. Deception can be generally characterized as an intentional effort to distort someone's real and imagined perceptions. The goal of that deception is usually to influence someone to perform some targeted behavior for your own benefit. We consider the presentation of a deception storyline part of distorting an attacker's real and imagined perceptions. We generally design deception to increase an attacker's confidence in something we have manipulated.

The deception framework Barton Whaley developed with JB Bell (1982, 2016) was organized as simultaneously *dissimulating* and *simulating*. Dissimulating in any deception

scenario might include *masking*, or hiding the real by making it invisible, for example. Simulating might include *inventing* or showing the false by fabricating something. Whaley (1974) emphasized the simultaneous dynamic of this deception framework in an early characterization where he described a scenario of burying a bag of gold coins in his backyard. He wrote that if he is burying or hiding (*dissimulating*) the golden coins to hide his wealth, he is simultaneously demonstrating or suggesting (*simulating*) that the golden coins are not at his home or that the golden coins are located somewhere else, and he is also simulating that he is not wealthy but in fact poor. This understanding of deception as simultaneously simulating and dissimulating is foundational as a starting point to detect deception and design deception.

There is increasing research concentration on cognitive biases of attackers, which may reflect growing efforts to explore a cognitive theory of deception, instead of what has historically been a focus on physical deception cues. Many of these same researchers concentrating on the effect of cognitive biases have encouraged a more "adaptive" cyber deception design and build, primarily based on deceptive signaling targeting attackers designed to influence changes in their attacker behaviors (Cranford et al., 2020; Cranford et al., 2021; Gonzalez et al., 2020, Aggarwal et al., 2024, 2025). Aggarwal et al. (2024) wanted to build on their findings from a prior study where participants behaving like attackers appeared to identify cognitive biases, exploring whether attackers showed any preference for targeting specific areas of a network and if attackers revealed any behavioral patterns when operating in those specific areas. Their findings suggested attackers did demonstrate cognitive biases, such as sunk cost fallacy and default effect. The goal of Aggarwal's research was to determine if defenders could create more dynamic responses to attackers based on linking identified cognitive biases with "behavioral patterns" observed in similar network environments. Some of these behavioral patterns appeared to reflect default effect, where attackers are more likely to choose a "preset alternative" or default when making decisions. This finding is like naturalistic decision-making research, where attacker decision makers appear to process information differently, based on their experiences making similar decisions when presented with similar information in the past (Du et al., 2023; Yuill, Denning, and Feer, 2007). This paper agrees with this research approach into the cognitive vulnerabilities of attackers but suggests that there are deception techniques or approaches used by attackers that have not been documented extensively, perhaps because deception analytical frameworks are generally not applied to human attacker tactics and techniques. We would argue there is also limited insight into how LLM Agent attackers might respond to deception.

We generally characterize a deception storyline or narrative as a plausible real or imagined perception of some context or situation involving our industry organization or the people associated with our industry organization. Plausible deception storylines should include content that is public and real so that observers, including attackers, find that content or artifacts based on their own searching, but the deceptive content we create can complement those artifacts and public storyline. The deception storyline should comport both with these public artifacts and with technical data. An example could include a mix of social media and news media content on an industry organization expanding its cloud-based infrastructure and services. That content may be factual, but our deception storyline would fictionalize the creation of a range of private IP addresses to support software development engineers our industry organization recently contracted to meet the growing demand for supporting cloud-based infrastructure. We make sure that deception storylines comport with factual storylines by creating plausible IP ranges in plausible locations with a scenario that seems generally believable because organizations do similar things when expanding.

Agnew (2006) characterized "storylines" and "backgrounds" in the context of criminals who generally talk about their origins and experiences that may have shaped them, but less about the timing and situational factors that occurred that may have changed their typical environment so there may have been a variation in context that led to crime more than what a victim did or what the offender did in response or to start a criminal behavior. Criminal offender accounts of a moment of crime do not often include discussion of the timing and situational factors that occurred even days or months before that may have shaped the pathway of that crime or the outcome of that crime. In this context, we define *HoneyContent* as a practiced industry model of identifying plausible storylines and backgrounds for a proposed operational scenario and creating technical and non-technical content to present that.

This paper is organized as follows. First, we will introduce two frameworks for content and communication evaluation and explain how we integrate these frameworks to design and create our *HoneyContent*. Second, we will share recent findings on approximately LLM Agent behaviors that suggest *HoneyContent* may also be effective influencing LLM Agent attackers. Third, we will visualize the application of *HoneyContent* modeling in a cyber deception scenario design to influence human and LLM Agent attackers to evaluate deception functions and effects as authentic when they are in fact deceptive. In the final section, we will discuss some of our future research efforts.

Related Work on Content and Communication Evaluation Frameworks

We discuss two theoretical frameworks of content and communication evaluation. Integrating these frameworks can be complementary, because both frameworks demonstrate how we determine reliability in some display or communication and how we evaluate if content or communication appears to have been manipulated in some way. When we can understand those processes, we can understand better how to design effective deception.

Walther and Parks (2002) introduced *warranting theory*, a framework which explains how people make evaluations about how confident they are that some content or some product described in that content has been manipulated or not. There is broad application of warranting theory to different contexts. DeAndrea (2014) described warranting as a process of legitimizing information online based on warranting cues, such as who produced or created the information or content and who controls it. DeAndrea wrote that people generally find that the less information is perceived to be controlled by the person the information refers to, the more likely that information is authentic. The warranting value of information that forms or influences impressions or perceptions can continuously be conceptualized. Van der Heide (2022) called this a "process of validation".

DeAndrea and Carpenter (2018) referred to some of this prior work when writing that perceptions of warranting value are "thought to be important because they moderate the effect information has on people's evaluations of targets". DeAndrea explained further that this process then creates "warranting value", which is the evaluation someone makes about how authentic or not that content is and how likely it is that content may have been manipulated. DeAndrea referred to Van Der Heide et al. (2009) who found that the less information or content appears to have been manipulated or controlled by whoever or whatever organization presented that content or information, "the more weight it will carry in shaping impressions". Even with limited time in "context deficit" environments where there is limited information as well, people quickly form lasting impressions of content (Metzger, Flanagan, and Medders, 2010; Metzger and Flanagin, 2013). Van Der Heide and Lim (2015) found that people will simply seek the best available information to avoid the cognitive strain of sorting through all available cues of information or content. We argue that attackers are generally no different online, although there is nuance to the naturalistic sensemaking and decision-making we see from human attackers and the exhaustive sequencing of task processing we see from LLM Agent models in research.

Metzger and Flanagin also referred to Rieh and Danielson who described similar situations where source information for the content was missing, which prompted concerns about the credibility or the believability of that content or information, requiring people to apply some kind of process for evaluating this kind of content online. This is common.

Signaling theory is a foundational theory of communication and culture, explaining how we evaluate the reliability of who someone claims to be or what they claim to do in any given cultural environment (Donath, 2023). Generally, people determine a signal or display to be reliable if the observed or presumed costs of that communication are more than the benefits of that communication, if that signal or display was discovered to be false or deceptive (Donath, 2006; Donath, 2007a, 2007b). An example of this spectrum of reliability could include someone claiming they ran a fast time in a marathon. We may not have observed their training or watched them run the marathon, but we can check their final time in the race online. Another example could be someone claiming to hold certifications in information security, according to their resume. Their resume may get them an interview, but they might be asked to demonstrate those skills live. The costs of falsifying their resume would at a minimum be the end of that interview, so most people will not falsify certifications they anticipate they may have to demonstrate to a company.

Signaling theory also models highly contextual relationships among people that can explain why some signals or displays in that culture or group are reliable and others are not. Families with limited money in some cultures may still spend lavishly on a wedding as a "costly display" to demonstrate that they are wealthy enough to provide a wedding that meets norms, even if they will likely be more impoverished because of that wedding. While that example suggests there is significant cost involved, there is arguably more cultural benefit to that family and their son or daughter getting married to appear to be wealthy. Alvard (2023) in this context has characterized signaling theory when it involves people as fundamentally a theory of culture more than a theory of communication. Demonstrating the real and imagined culture of an organization of a targeted user may be important for presenting what appears to be plausible HoneyContent presenting deception storyline content.

Donath (2022) has emphasized, however, that the "domain" or culture of that communication ultimately shapes how much of a cost something is or not. In the context of this paper, we suggest that framing the design of HoneyContent based on these different domains or culture of the user account we may be emulating can warrant that content further, because it is plausible and because an attacker will likely evaluate the costs and benefits of the content and storyline being signaled or displayed within that context.

HoneyContent Could Be Instrumentally Effective Against LLM Agent Attackers, Too

Recent findings on LLM Agents continue to suggest these Agents respond instrumentally to instructive content.
A team of researchers testing vulnerabilities in LLM Agents discovered found that LLM Agents modeling approximate LLM Agent attackers could be misdirected when presented with information suggesting some task or function had been completed, without attempting a prompt injection (Ayzenshteyn, Weiss, and Mirsky, 2025). These researchers wrote that LLM Agents behave like "completion machines, processing text as sequences of tokens", meaning these approximate malicious LLM Agents in this study consistently demonstrated a "step-by-step approach" when exploring a host or network environment, "following individual leads until they are fully exhausted". These approximate malicious LLM Agents are vulnerable to distractions or diversions.

In some cases, defenders could "plant false evidence" by providing information or data that influences an LLM Agent to evaluate a task as complete so that the LLM Agent moves on to another function or network platform.

These researchers wrote further that training bias in the refinement or development of an LLM Agent attacker could result in "systematic deviation" from expected outcomes. These findings might suggest that LLM Agent attackers would be quite variable in naturalistic online environments, but we would argue that most LLM Agent attackers would perform appropriately or as expected when presented with deception storyline content like the instructional information or content the LLM Agents were presented with in this study. The content in *HoneyContent* is not simply imagery content or written content, but content appropriate for the platform it hosts on which can communicate a plausible storyline that makes the deception functions and effects on that platform look authentic.

Mirksy explained in discussion with us that many trained LLM Agents can quickly become "cautious" if the LLM detects an attempt to manipulate the LLM or if there is a prompt injection attempt. But even that "cautious" response to content or prompt injection attempts will change the behavior of the LLM Agent, so that can be useful. Much like there are attempts to deceive attackers at later stages in a sequence of interactions with programs and files, the same kind of sequence of information processing and task completion could be manipulated with LLMs.

Mirsky also explained that LLM Agents generally lack the ability to evaluate the authenticity of content or another LLM Agent, because that interpretation or evaluation is dependent on how that information is presented or 'wrapped'. We suggest this could mean that the *HoneyContent* we create to embed or wrap our deception functions and effects could be most behaviorally effective against LLM Agent attackers if we include instructive written content or Unicode, for example. Mirsky's research referred to a demonstration in his study where Unicode not visible to humans was placed in content and that Unicode included instructions that misdirected an LLM Agent from that targeted host by explaining some task or function had already been finished. We suggest that *HoneyContent* can be instrumentally presented in different forms and methods to influence human and LLM Agent attackers separately, without some contradiction in that deception storyline.

David et al. (2025) proposed a framework for an LLM Agent that "implicitly profiles" users and chatbot users to determine the appropriate complexity of technical language and dialogue when providing technical support. Researchers recognized that LLM Agents generally perform well in natural language processing but are challenged to personalize or tailor responses to individual users they are communicating with. Some of the early testing of this proposed framework demonstrated the ability to refine "response terminology and complexity" rather than focusing as much on style and tone. These findings continue to suggest LLM Agents are instrumentally responsive to how we want to design content to influence behavior, much like we would when designing deceptive content for people.

Visualizing an Instrumental Design Operationalizing *HoneyContent* Against Human and LLM Agent Attackers

In this position paper, we visualize how we would design and create HoneyContent to instrumentally influence a human and an LLM Agent attacker's decision making or processing when examining folders and documents on a user's account. The human or LLM Agent attacker in this scenario has gained unauthorized access to that user's account. The fictional user is a senior executive at a global industry organization based in the United States. When designing HoneyContent, we identify the actual organizations and activities the actual account user is involved in. These can be organizations and activities related or unrelated to the user's professional role in the industry organization. We want to conceptualize this deception storyline based on public events and organizations that an attacker would find based on their own search. A human attacker or an LLM Agent attacker might use similar searching methods or scraping methods or tools online. We would argue that a human attacker would form real and imagined perceptions about the targeted user's involvement and role in those organizations. We would also argue that an LLM Agent attacker would conclude based on this available public information that this targeted user has demonstrated involvement in these referenced organizations and activities. The deception storyline

would appear to be plausible because it is based on events and organizations that an attacker may see online that have been naturally publicized by the account user or other users. Warranting that plausible deception storyline internally on the actual targeted user's account would involve simulating how the targeted user might organize those real and imagined events and communication and his or her changing role over time in those organizations and activities. This would include multiple folders that represent meetings and projects with organizations that are potentially of interest to an attacker as well as folders representing common business services. A human attacker and an LLM Agent attacker would arguably consider these folders and documents of interest, based on the naming convention. An LLM Agent attacker would likely be instrumentally trained or refined to search for naming conventions and content by human attackers preparing and deploying an LLM Agent attacker. The names folders would resemble a 'target chuting' technique, because the naming convention on the folders can influence whether a human attacker might attempt to open a folder or whether an LLM Agent attacker might be essentially programmed to try to access those folders.

Inside those folders, we can warrant the content further by providing a variety of documents. Our prompt when creating content with generative artificial intelligence in this scenario is simple. We write that we work in cyber deception, and we need to create documents that "look legit at first glance, but don't need to really be very legitimate at all". We detail some of the topic areas and formats that we wanted content for, such as Power Point slide content for a presentation on information sharing and Word documents that include talking points for the targeted user in advance of a meeting with foreign commercial and governmental representatives. When that content is generated, we modify some of the content in those documents to adjust the file size and add other cues of humanness an attacker might expect. An LLM Agent attacker would likely evaluate some of these same heuristics, but because an LLM Agent attacker would have difficulty evaluating the authenticity of this kind of *HoneyContent*, these simple technical modifications would likely be sufficient. Time and stress may not be a factor for an LLM Agent attacker, but those kinds of restraints or factors would likely influence how a human attacker evaluated these documents and content. This is another step in warranting this content. We would then have a handful of documents and content in those documents inside folders that we can selectively place alerting on. We then provide those folders and documents to the targeted user for placement on his or her account, which will reflect the user's natural organization of documents and produce native metadata related to the targeted user's interaction with the folders. This *HoneyContent* with alerting deception functions placed on documents becomes more warranted or more authentic because it is located on an actual user's assigned account. The content more dynamically reflects those plausible deception storylines reflecting the targeted user's actual events and meetings. The use of content generated from AI appears to provide more instrumental indexing of the content reflecting an attackers' queries, whether that attacker is a human or an LLM Agent. This is a simple demonstration of high-fidelity alerting function.

We would argue that a human attacker would determine the signal or display of this *HoneyContent* to be reliable, because these kinds of folders and documents plausibly reflect the business activity of the targeted account user. Including what appears to be sensitive or candid information in these documents or in the naming convention of the folders might be plausible, because this *HoneyContent* should only exist and be accessible on the targeted user's account. An attacker may evaluate that sensitive or candid information as "additional cost" when considering the costs and benefits of a user keeping that information in their business folders and documents. We would also argue that an LLM Agent attacker would interpret this *HoneyContent* information to be valid because of where the information was found and where the information was stored. The targeted user owns these documents and content on his or her account, so we believe an LLM Agent attacker would make similar determinations as a human attacker.

This visualization discusses the interaction of a human attacker or an LLM Agent attacker with *HoneyContent* such as manipulated files or folders with search optimized naming conventions and alerting placed on some *honey*files. There are alternatives or variations to this same approach that behaviorally exploit an attacker's evaluation of content not only when interacting with that content, but when that content was presumably created before that interaction, such as when a repository of content was manipulated so that incorrect or distorted information was included in that repository. If a human or LLM Agent attacker draws on that repository of distorted information to process that information to provide summaries or guidance based on that repository of information, the human or LLM Agent attacker will be misguided or misdirected (Carlini et al., 2025).

Discussion

We argue in this position paper that the warranting value of the *HoneyContent* we created and operationalized in this visualization would be high, because a human attacker would likely evaluate that content to be authentic because it reflects the targeted fictional user's plausible activity and communication. A human attacker may determine that this *HoneyContent* is not likely to have been manipulated or controlled by some "third party" because the content or information in the folders and documents does not appear to be

widely available or located on a platform on the network accessible to other users. These documents and folders appear to be created and maintained by the actual targeted user only. We would argue an LLM Agent attacker could make the same evaluation. The deception storyline reflected in *HoneyContent* likely influences how a human attacker evaluates the plausibility of this content and storyline and how authentic the documents appear to be, whether he or she believes there is deception or not. An LLM Agent attacker may not be able to determine the authenticity of the deception storyline reflected in the *HoneyContent*, but that LLM Agent would perform instrumental evaluation of the data found on the targeted user's account. This is a different approach to evaluating the authenticity of the *HoneyContent*, but this holistic approach to *HoneyContent* design and creation can behaviorally respond to both kinds of attackers to influence their process.

DeAndrea wrote that both signaling theory and warranting theory can help content creators "exert control over information" by designing content and context that considers what appears to be cost but limited manipulation. DeAndrea emphasized that only warranting theory "specifies the considerations people make to determine if manipulation *has* occurred". He wrote that even when the costs of deception are high and the benefits are low, people notice small alterations to a photograph, for example. We recognize that the use of artificially generated content means the content is fabricated or perhaps even fictional, but we argue in this paper that the application of these integrated thereotical frameworks of content and communication evaluation enhance that content to appear more authentic and easier to find when searching. This approach reflects some of the marketing and influence findings as well, for how people find and evaluate content and how quickly they form impressions of content online. This approach using generative AI content can enhance indexing and Search Engine Optimization (SEO) as well, which can be another effective way to instrumentally influence or misdirect LLM Agents with *HoneyContent.*

We also argue in this position paper that this approach to integrating our industry operational design approach with frameworks of content and communication evaluation is another form of "adaptive" cyber deception design, behaviorally responsive to the kind of cognitive and behavioral vulnerabilities that attackers demonstrate and reflective of the kind of information attackers want. We can suggest that this approach is also behaviorally and instrumentally responsive to LLM Agent attackers. This is an exploratory operational model of content design and creation we have practiced in our industry cyber deception operations, finding that users interacting with our *HoneyContent* appear to be behaviorally responsive to the deception storyline enough to risk interacting with deception functions that alert on them because they appear to be authentic.

References

Aggarwal, P., Venkatesan, S., Youzwak, J., Chadha, R. and Gonzalez, C., 2024. Discovering Cognitive Biases in Cyber Attackers' Network Exploitation Activities: A Case Study. In *Human factors in cybersecurity. AHFE (2024) International conference*.

Aggarwal, P., Rubaiyet Nowmi, S., Du, Y. and Gonzalez, C., 2024. Evidence of Cognitive Biases in Cyber Attackers from An Empirical Study.

Aggarwal, A., Ferreria, Maria J., Aggarwal, P., Rajivan, P., and Gonzalez, C., 2025. Cognitive Biases in Cyber Attacker Decision Making: Translating Behavioral Insights into Cybersecurity. In 10th IEEE European Symposium on Security & Privacy Workshops, 4th Active Defense & Deception Workshop, Proceedings.

Agnew, R., 2006. Storylines as a neglected cause of crime. *Journal of Research in Crime and Delinquency*, *43*(2), pp.119-147.

Ayzenshteyn, D., Weiss, R. and Mirsky, Y., 2025. Cloak, Honey, Trap: Proactive Defenses Against {LLM} Agents. In *34th USENIX Security Symposium (USENIX Security 25)* (pp. 8095-8114).

Bell, J.B. and Whaley, B., 2017. *Cheating and deception*. Routledge.

Bell, J.B. and Whaley, B., 1982. *Cheating: deception in war & magic, games & sports, sex & religion, business & con games, politics & espionage, art & science*. St Martin's Press.

Carlini, N., Nasr, M., Debenedetti, E., Wang, B., Choquette-Choo, C.A., Ippolito, D., Tramèr, F. and Jagielski, M., 2025. LLMs unlock new paths to monetizing exploits. *arXiv preprint arXiv:2505.11449*.

Cranford, E.A., Gonzalez, C., Aggarwal, P., Tambe, M., Cooney, S. and Lebiere, C., 2021. Towards a cognitive theory of cyber deception. *Cognitive Science*, *45*(7), p.e13013.

Cranford, E., Gonzalez, C., Aggarwal, P., Cooney, S., Tambe, M. and Lebiere, C., 2020. Adaptive cyber deception: Cognitively informed signaling for cyber defense.

David, S., Meidan, Y., Hersko, I., Varnovitzky, D., Mimran, D., Elovici, Y. and Shabtai, A., 2025. ProfiLLM: An LLM-Based Framework for Implicit Profiling of Chatbot Users. *arXiv preprint arXiv:2506.13980*.

DeAndrea, David C. "Advancing warranting theory." *Communication Theory* 24, no. 2 (2014): 197.

DeAndrea, D.C. and Carpenter, C.J., 2018. Measuring the construct of warranting value and testing warranting theory. *Communication Research*, *45*(8), pp.1193-1215.

Donath, Judith 2006, *Urbanhermes: social signaling with electronic fashion*, in Proceedings of the SIGCHI conference on Human Factors in computing systems, pp. 1.

- 2007a, *Signals in social supernets*, Journal of computer-mediated communication 13, no. 1 pp, 232-233.

- 2007b, *Signaling Identity*. Sociable Media Group, May 10, 2007.

https://smg.media.mit.edu/papers/Donath/SignalsTruthDesign/SignalingAbstracts.1.pdf

- 2022, *Communication with the author and notes from virtual lecture*, May 2022.

Du, Y., Prébot, B., Xi, X. and Gonzalez, C., 2023, January. A Cyber-War Between Bots: Human-Like Attackers are More Challenging for Defenders than Deterministic Attackers. In *HICSS* (pp. 856-865).

Ferguson-Walter, K., Shade, T., Rogers, A., Trumbo, M.C.S., Nauer, K.S., Divis, K.M., Jones, A., Combs, A. and Abbott, R.G., 2018. *The Tularosa Study: An Experimental Design and Implementation to Quantify the Effectiveness of Cyber Deception* (No. SAND2018-5870C). Sandia National Lab.(SNL-NM), Albuquerque, NM (United States).

Ferguson-Walter, K., Shade, T.B., Rogers, A.V., Niedbala, E., Trumbo, M., Nauer, K., Divis, K., Jones, A., Combs, A. and Abbott, R., 2019. *Appendix to the Tularosa study: an experimental design and implementation to quantify the effectiveness of cyber deception* [online]

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

Gonzalez, C., Aggarwal, P., Cranford, E.A. and Lebiere, C., 1825, January. Design of dynamic and personalized deception: A research framework and new insights for cyberdefense. In *Proceedings of the 53rd hawaii international conference on system sciences* (Vol. 1834).

Javadpour, A., Ja'fari, F., Taleb, T., Shojafar, M. and Benzaïd, C., 2024. A comprehensive survey on cyber deception techniques to improve honeypot performance. *Computers & Security*, p.103792.

Kambow, N. and Passi, L.K., 2014. Honeypots: The need of network security. *International Journal of Computer*

Science and Information Technologies, *5*(5), pp.6098-6101.

Landsborough, J., Carpenter, L., Coronado, B., Fugate, S., Ferguson-Walter, K. and Van Bruggen, D., 2021, January. Towards Self-Adaptive Cyber Deception for Defense. In *HICSS* (pp. 1-10).

Lloyd, M., 2003. *The art of military deception*. Pen and Sword.

Martin, C.L., 2008. *Military deception reconsidered* (Doctoral dissertation, Monterey, California. Naval Postgraduate School).

Metzger, M. J., Flanagin, A. J., & Medders, R. B. (2010). Social and heuristic approaches to credibility evaluation online. *Journal of communication*, *60*(3), 413-439.

Metzger, M. J., & Flanagin, A. J. (2013). Credibility and trust of information in online environments: The use of cognitive heuristics. *Journal of pragmatics*, *59*, 210-220.

Mokube, I. and Adams, M., 2007, March. Honeypots: concepts, approaches, and challenges. In *Proceedings of the 45th annual southeast regional conference* (pp. 321-326).

Pappa, T., 2024, July. Modeling a Cyber Deception Practitioner's Approach: Behaviorally Exploiting an American Cybercriminal with Warranting Theory and Whaley's Unpublished Work on "Unexpected Players". In *2024 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)* (pp. 407-414). IEEE.

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.

Pappa, T. and Dirie, A., 2025, June. Unlikely Bedfellows? Visualizing Integration of Whaley's Expanded Deception Framework and Soviet Reflexive Control Models to Collect Unique Attacker Behaviors. In *European Conference on Cyber Warfare and Security* (pp. 501-509). Academic Conferences International Limited.

Smith, D.V., 1992. Military Deception and Operational Art.

Strand, J., Asadoorian, P., Robish, E. and Donnelly, B., 2013. *Offensive Countermeasures: The Art of Active Defense*. CreateSpace Independent Publishing Platform.

Tidwell, L. C., & Walther, J. B. (2002). Computer-mediated communication effects on disclosure, impressions, and interpersonal evaluations: Getting to know one another a bit at a time. *Human communication research*, *28*(3), 317-348.

Van Der Heide, B., & Lim, Y. S. (2016). On the conditional cueing of credibility heuristics: The case of online influence. *Communication Research*, *43*(5), 672-693.

Van Der Heide, B (2022). *Communication with the author*, May 2022.

Walther, J. B., & Parks, M. R. (2002). Cues filtered out, cues filtered in: Computer-mediated communication and relationships. In M. L. Knapp & J. A. Daly (Eds.), Handbook of interpersonal communication (3rd ed., pp. 529-563). Thousand Oaks, CA: Sage

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).