

Agentic AI systems Conducting Social Engineering Attacks

- Julian Neylan

EU CyberNet Expert Series
No 7
2026

a b o u t
**Julian
Neylan**



Julian Neylan is the Dissemination and Defense Lead at the Disarm Foundation and Training Programme Lead with Alliance 4 Europe. His research interests are in disinformation, AI, cybersecurity and public health.

Julian has been part of the EU CyberNet Expert Pool since August 2025.

summary

This article highlights how Agentic AI systems can be used in various steps of the social engineering kill chain. It highlights measures and broader policy recommendations of actions that can be taken to mitigate the potential harms from Agentic AI automated attacks.



Introduction

Agentic AI has the potential to make social engineering attacks more personalized and scalable [1], enabling attackers to target large numbers of individuals while maintaining the appearance of individualised communication. By Agentic AI we refer to autonomous systems that can plan and execute multi-step tasks with limited human supervision, often by orchestrating multiple AI models and external tools.

Social engineering used to be constrained by human labor: researching targets, crafting believable pretexts and sustaining back-and-forth engagement took time and skill. Agentic AI changes that constraint by enabling software systems that can plan, execute and iterate multi-step tasks with minimal human direction. We've already started seeing AI used in attack flows as Anthropic detailed an espionage hacking campaign that operated 80-90% [2] automatically with minimal human intervention. We know open source tools like the Social Engineering Toolkit or Gophish can be utilised by Agentic AI to conduct social engineering attacks at scale. Additionally, Agentic AI usage can be facilitated by open source tools like Langchain. Open-source reconnaissance tools can automatically collect emails, employee names, subdomains and other publicly available information about organisations. When integrated into agentic systems, these tools allow attackers to build detailed profiles of targets at scale.

In this blog we will overview how Agentic AI systems can take advantage of open source tools along all of the stages of the social engineering kill chain. We then provide recommendations for how to mitigate this problem.

Stages of the social engineering kill chain [3]

Reconnaissance

The first stage of most social engineering campaigns involves collecting intelligence

about the target. Attackers gather information about individuals, organisations, roles, communication patterns and potential vulnerabilities. This information enables attackers to design convincing narratives and identify leverage points within an organisation.

Agentic AI could dramatically expand the scope and efficiency of reconnaissance activities. Autonomous agents can aggregate information from publicly available sources (e.g. including professional networking sites, social media, company websites and leaked data repositories) and synthesize it into structured profiles of targets.

Separate open-source reconnaissance tools (e.g., theHarvester and Recon-ng) can accelerate collection of publicly available emails, names, subdomains and other OSINT needed for collecting data on potential targets as well as their networks.

Preparation and Weaponization of Information

In this stage, the attacker uses the information gathered during the reconnaissance to craft the attack scenario that would best appeal to the victim's psychological profile together with the necessary resources and infrastructure to conduct the attack.

Infrastructure can be in the form of fake identities and impersonations of legitimate people. Fake account creation can be accelerated using Agentic AI systems that can use public tools like this person does not exist or their own image generation capacity to create images and LLMs to write bios. LLMs can be given specific identities and personas often through open source repositories.

They also may choose to impersonate existing entities known to be trusted by the victim. Major fraud incidents show how synthetic media can be operationalised in social engineering. The Financial Times reports that engineering firm Arup lost about US\$25

million after a Hong Kong employee was deceived during a video conference involving deepfaked senior colleagues [4]. Open source tools like Open Voice can be used to clone voices.

Initial Contact

Initial contact can be done in the form of email, DMs, phone calls, texts and social media outreach. Text phishing and spear phishing remain foundational and LLMs can improve the speed and quality of text generation. In one experimental study [5], participants often ranked AI-generated spear-phishing messages as more persuasive than those written by humans, highlighting the potential effectiveness of AI-assisted attacks. In other formats voice audio can be used to conduct phone or video calls. Open source tools like Rasa or Botpress could be used to create conversational Agentic AI.

Establishing a Rapport

AI-enhanced social engineering campaigns benefit from advanced targeting and personalization [6], allowing attackers to tailor messages to specific individuals or roles. Autonomous conversational agents can converse with victims in real time to develop a relationship. Large language models can generate emails, chat messages or scripts that mimic professional communication styles reducing the linguistic errors that historically made phishing messages easier to detect. Attackers can use AI tools to tailor messages to their target, as they can exploit personal details, behavioral patterns, and communication styles [7].

Exploitation

This is where the victim performs the specific action intended by the attacker, this can be disclosure of credentials or sensitive information, malware deployment, providing access to restricted data or infrastructure or access to other individuals or the transfer of funds [3]. The victim may allow exploitation

because they trust who they are speaking with due to the rapport they've developed, or they have been fooled further by technical means (e.g. by a mirror version of a website, which can be facilitated by agentic AI using open source tools like HTTrack, or by more traditional phishing).

Post-Exploitation

This is where the attacker decides to continue exploitation (gathering more information, exfiltrating the information or getting more access) or terminating the operation [3]. This is likely the point where a human in the loop would help the Agentic AI system determine whether to stop or whether to keep going. The agentic AI system may also continue indefinitely if it has not achieved the specific aims that were assigned to it.

How to mitigate the social engineering capabilities of Agentic AI systems

Steps can be taken to help mitigate the attack surface available to agentic AI systems. This can include:

- Limit the amount of publicly available information available online which can be used by Agentic AI systems. This can be either text based information or images, audio or video of the target. Organisations need to limit employee details on public pages to protect these individuals and review what information they put on social media.
- More thorough authentication needs to be required for important interactions such as financial transactions. This might be by having more individuals be required to be involved to complete transactions.
- If attackers utilize web crawlers, changing the robots.txt file of a website may help if they use a crawler that doesn't ignore the robots.txt. Additionally, server controls can be used to block agentic scrapping, vendor tools like Cloudflare bot detection or specialised agent detection tools.

- Alerting people of the advanced personalization capabilities of agentic AI so they are less likely to respond to attempts to target them. Right now, individuals need to be aware that the presence of not only themselves but also their closest connections can be utilised in highly advanced social engineering attacks. [Audio](#) and [video](#) deepfakes cannot be reliably detected by people. Normal cybersecurity is meant for large organisations and institutions, however the barrier to conducting advanced campaigns has been lowered to the extent that it is now easier and more worthwhile for actors to target individuals.
- Developers should consider safeguards that reduce the risk of abuse, such as documentation on ethical use, rate-limiting features or abuse monitoring. Open source is open to truly anything. Tooling designed for defensive purposes

like penetration testing or simply for fun can be used to create sophisticated campaigns. Those publishing repositories should consider how their software may be used in social engineering campaigns.

- AI companies need to ensure that there are proper guardrails against their products being used in advanced social engineering attacks.

Agentic AI lowers the cost and skill barrier required to conduct sophisticated social engineering campaigns. By automating reconnaissance, message generation and interaction with victims, attackers can scale operations that once required significant human effort. As these capabilities become more accessible, organisations and policymakers must adapt their defenses to a threat landscape where deception is increasingly automated.



References

- [1] CrowdStrike (2026) Social Engineering and AI <https://www.crowdstrike.com/en-us/cybersecurity-101/social-engineering/ai-social-engineering/>
- [2] Anthropic. (2025, November 13). Disrupting the first reported AI-orchestrated cyber espionage campaign. <https://www.anthropic.com/news/disrupting-AI-espionage>
- [3] Nowakowski, W. (2025). Social Engineering Analysis Framework: A Comprehensive Playbook for Human Hacking. IEEE Access, 13, 18827-18849.
- [4] Financial Times. (2024). Arup lost \$25mn in Hong Kong deepfake video conference scam. <https://www.ft.com/content/b977e8d4-664c-4ae4-8a8e-eb93bdf785ea>
- [5] Francia, J., Hansen, D., Schooley, B., Taylor, M., Murray, S., & Snow, G. (2024). Assessing AI vs human-authored spear phishing SMS attacks: An empirical study. arXiv preprint arXiv:2406.13049.
- [6] Schmitt, M., & Flechais, I. (2024). Digital deception: Generative artificial intelligence in social engineering and phishing. Artificial Intelligence Review, 57(12), 324.
- [7] Dzuba, C. (2025). Artificial intelligence in social engineering: a literature review through the lens of routine activity theory. Issues in Information Systems, 26(2).

Agentic AI systems Conducting Social Engineering Attacks
Julian Neylan

EU CyberNet Expert Series, No 7, 2026

© EU CyberNet 2026



Funded by
the European Union

Views and opinions expressed are those of the author only and do not necessarily reflect those of the European Union or the EU CyberNet. Neither the European Union nor EU CyberNet can be held responsible for them.