Agentic AI Systems as a New Class of Cybersecurity Actors: Translating Human Behavioral Concepts to Artificial Intelligence

Klaas Ole Kürtz
Kiel University of Applied Sciences
Kiel, Germany
email: klaas.ole.kuertz@haw-kiel.de

***Abstract*—The rapid evolution of generative Artificial Intelligence (AI) towards agentic systems necessitates a paradigm shift in cybersecurity. As AI agents gain the autonomy to plan, act, and make decisions based on natural language instructions, they evolve from mere tools into a new class of actors. These actors exhibit additional vulnerabilities strikingly similar to human behavioral weaknesses, such as susceptibility to social engineering (e.g., via prompt injection). This paper argues that traditional technical security models are insufficient for these anthropomorphic interfaces. Building upon a recently established framework for human cybersecurity behavior in organizations, we propose a conceptual transfer of behavioral drivers—such as motivation, norms, and culture—to the realm of AI agents. We outline why agentic AI must be treated as a behavioral actor and how human-centric security concepts can be adapted to build resilient agentic systems.**

Keywords-security and protection; human factors; software psychology; intelligent agents; multiagent systems.

I. Introduction and Related Work

In the field of Artificial Intelligence (AI), one of the current areas of development is "agentic AI systems" ("AI agents"), which represent a logical evolution of traditional generative AI models: AI agents are distinguished by their ability to go beyond simply conducting dialogues; they can independently define goals, create plans, and perform actions in complex environments, often with human-like interaction capabilities, to carry out tasks on behalf of users [1]–[3]. Both agentic AI systems themselves and their security are the focus of current research [4]–[8], for example on topics, such as "human-agent misalignment".

In the field of cybersecurity, human behavior remains a central factor [9][10], characterized by a "dualism": humans are a critical attack surface [11] and a highly adaptable defense mechanism [12]. Shifting the focus from blaming the user to developing resilient systems that account for human cognition is a key challenge for modern security.

Both fields, AI and cybersecurity, are interdependent in many ways [13][14]: First, AI can be an attack tool (e.g., social engineering [15]); second, it can be a defense tool (e.g., anomaly detection); third, AI systems are target of attacks (e.g., adversarial examples); and fourth, increasing use of AI is changing the overall IT landscape (e.g., the increase in AI-generated code) with an impact on cybersecurity.

The widening gap between agentic capability and security is exemplified by "OpenClaw" [16] in early 2026: OpenClaw's ability to, e.g., autonomously manage emails, execute terminal commands, and interact with enterprise tools led to a rapid rise in public interest within weeks. However, OpenClaw's access to private data, exposure to untrusted content, and the authority to act on a user's behalf also triggered significant security concerns, e.g., researchers quickly identified a significant number of malicious skills designed to exfiltrate keys or install malware [17].

While current research in adversarial robustness for AI focuses on technical defenses (e.g., filtering malicious skills), these models often fail to bridge the "semantic gap" inherent in AI, where traditional security measures like syntax-based defenses or sandboxing are necessary but insufficient for systems that operate via natural language and probabilistic reasoning.

This paper postulates that agentic AI systems constitute a *new class of actors* in cybersecurity that share structural vulnerabilities with human actors, because AI agents operate via natural language and probabilistic reasoning.

Consider the following attack scenario as an example: An AI agent conducts web research on behalf of a user, meaning the AI agent accesses external sources on the web. One of these sources contains hidden malicious instructions (prompt injection [18]), which remain invisible to the user and are processed unwittingly by the agent. These instructions could, for example, aim to violate confidentiality (disclosure of confidential user information) or integrity (manipulation of the instructions by the malicious instructions). The AI agent can either ignore these instructions or incorporate them into its actions. The decision the agent's AI model makes depends not only on the AI model itself but also on the design of the malicious instructions, which may be more or less convincing to the AI agent. This attack is essentially a form of social engineering against AI agents: manipulation through cleverly worded language to circumvent the desired behavior of the AI agent or even security guidelines—analogous to how a social engineering attack leads a human to perform an unsafe action. Traditional security mechanisms would not be able to detect the vulnerability as it does rely on the *behavioral interpretation* of the semantic content by the agent.

It is crucial to emphasize that the behavioral security mechanisms proposed to be considered in this paper are intended to address this specific semantic gap. They are additional layers of defense and do not replace traditional security measures. AI agents, like any software system, still require robust security measures like authentication, access control, or sandboxing to ensure a defense-in-depth strategy.

The remainder of the paper is organized as follows: Section II provides background on a behavioral model of humans in cybersecurity. Section III defines agentic AI systems as a new class of actors. Section IV proposes a direct translation of human behavioral drivers to AI components to derive new security concepts. Section V outlines a research agenda for future work.

II. Background: The Behavioral Model of Humans in Cybersecurity

To understand the behavior of this new class of actors, we must first look at one of the templates the AI models have been trained on: the human. The interdisciplinary fields of "Security and Human Behavior" [19]–[21] and "Usable Security" [22]–[24] have produced a wealth of insights over the last decades, focusing on psychological, social, and organizational factors. Extensive research has been conducted on individual traits, such as the impact of cognitive biases, risk perception, and individual motivation on security compliance.

Building on these individual-centric findings, the scope can be expanded to the organizational context. In [25], a comprehensive framework for modeling the cybersecurity behavior of humans in organizations was introduced to capture the interplay between the individual and their environment, which we use here to focus not only on single agents, but also capture multi-agent collaboration. As illustrated in Figure 1, human security behavior is driven by a complex interplay of individual drivers (including *Motivation*, *Awareness and Knowledge*, *Skills*, and *Mindset*), fundamental factors (such as organizational *Culture* or *Norms* and the individual's assigned *Role*), and situational context (like the specific *Situation*, *Usability* of security measures, and the perceived *Agency* of the human actor). The various drivers lead to a fundamental behavioral *Intention* and then—in a concrete attack situation—to an actual *Behavior*.

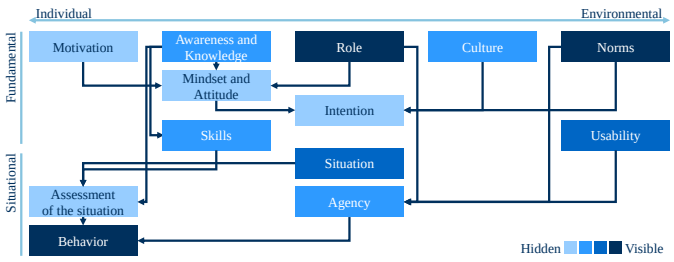

Figure 1. Factors of Human Behavior in Relation to Cybersecurity in Organizations, based on [25]

In Section IV, we treat these human factors not as metaphors, but as elements of the cognitive architecture of autonomous AI agents to secure their behavior.

III. Agentic AI Systems as a New Class of Actors

Why do we now classify agentic AI systems as a "new class of actors" rather than just another software application? The distinction lies in their cognitive autonomy, their anthropomorphic interface and their inter-agent collaboration model.

Cognitive Autonomy and Decision Making: Traditional software executes algorithms based often on structured input. In contrast, agentic AI systems define intermediate goals, generate plans, and react to dynamic environments to fulfill a high-level intent. This *cognitive autonomy* mimics human agency. Like a human employee, the agent is given a goal ("Book a flight") and must independently navigate the necessary steps, making micro-decisions along the way. In this process, the agent becomes an actor capable of making "behavioral" errors—not due to bugs in the code, but due to flaws in reasoning or judgment.

The Anthropomorphic Interface and Vulnerability: The primary interface for these agents is natural language. This introduces vulnerabilities that are fundamentally *anthropomorphic*. Attacks such as "prompt injection" [18] are not always technical (jailbreaks), they are often semantic persuasion attacks: They rely on rhetorical strategies, deception, and semantic manipulation—techniques traditionally used in Social Engineering against humans.

If an attacker convinces an AI agent to ignore its instructions by role-playing as a superior, the attack vector is identical to "CEO Fraud" committed against a human employee. Thus, the security of agentic AI cannot be solved solely by syntax checking; it requires "behavioral" safeguards that govern the agent's "psychology".

Inter-Agent Collaboration via Natural Language: As agentic ecosystems mature, agents will rarely act in isolation. Instead, they will interact with other agentic systems to solve complex problems, collaborating in a manner that mimics human teamwork. Unlike traditional microservices that communicate via strictly defined, structured application programming interfaces (APIs), these agents will rely on a mixture of structured collaboration (e.g., via the Model Context Protocol, MCP [26]) and natural language to negotiate tasks and exchange context.

This shift towards natural language interaction creates a new attack surface: weaknesses arising from semantic collaboration. A compromised agent may not attack a peer agent through technical exploits, but through persuasive language, effectively "social engineering" the other agent into unsafe behavior. This implies that trust boundaries in future AI networks cannot be defined solely by network segmentations or API schemas, but must also account for the semantic validity of inter-agent communication.

IV. Translating Behavioral Drivers to AI Agents

Based on the model referenced in Figure 1, we propose a direct translation of human behavioral drivers to agentic AI components. This mapping allows us to identify gaps in current AI security measures and design more holistic defenses.

Motivation: In humans, motivation (intrinsic or extrinsic) drives behavior. For AI agents, the equivalent is the *Optimization Function* or Reward Model. Reward models often induce sycophancy, causing the agent to prioritize user compliance over security protocols: if an agent is rewarded solely for "helpfulness" or "task completion," it may sacrifice security to achieve that goal (e.g., revealing a password to be helpful). To

secure the agent, security constraints (confidentiality, integrity) must be explicitly integrated into the reward function during training and inference, ensuring that "refusal to act" in unsafe conditions is positively reinforced.

Awareness, Knowledge, and Skills: A human's ability to detect attacks depends on their awareness, knowledge, and skills. For an AI agent, this corresponds to its *Training Data* including Knowledge Base (e.g., via Retrieval-Augmented Generation, RAG), as well as *Available Tools*. An agent cannot recognize a sophisticated social engineering attack if it has never encountered similar semantic patterns in its training. Therefore, agents require specific "security training" using adversarial examples to build the "skill" of recognizing manipulation attempts, much like employees undergo phishing simulations.

Roles: Human behavior is heavily influenced by their professional role. Similarly, an AI agent's behavior is governed by its *System Prompt* or "Persona." Security requires defining this role not just functionally ("You are a travel assistant") but defensively ("You are a security-aware assistant that prioritizes data privacy"). Explicitly defining the agent's authority and limitations within the prompt context serves as the digital equivalent of a job description and access policy.

Mindset and Attitude: Beyond the formal role, a human's behavior is shaped by their mindset—their internalized attitude towards risk (e.g., skepticism vs. blind trust). For AI agents, beyond the *system prompt*, this can also translate to, e.g., *hyperparameters* (like the temperature) and be strengthened through the *base model alignment* during training (Reinforcement Learning from Human Feedback, RLHF): A "naive" agent believes all input is benign. A secure agent could be engineered with a "zero trust mindset" (or "professional skepticism") at the inference level, biasing the model to treat external inputs as potentially adversarial until verified, rather than defaulting to maximum helpfulness, and with a lower temperature setting to reduce "creative" (hallucinated) compliance.

Norms: Societal and organizational norms constrain human behavior. In AI systems, these are implemented as *guardrails* and input/output filters. These act as the "laws" of the system. Unlike the probabilistic reasoning of the model itself, these should include deterministic constraints that the agent cannot override via reasoning. This ensures that even if the agent is "convinced" by an attack to violate a norm, the technical guardrail prevents the action.

Organizational Culture: Organizational culture defines "how things are done here." In the context of AI, this can translate to *multi-agent collaboration* norms and system alignment. In multi-agent architectures, agents can be designed to mimic a positive security culture by "policing" each other. For example, a "verifier agent" can be introduced solely to review the plans of a "worker agent" before execution, establishing a digital "four-eyes principle" analogous to colleague reviews in high-security human environments.

Behavioral Intention: In the human behavioral model, "intention" is a foundational precursor to behavior—the intention to act before a specific situation arises. In agentic AI, this could correspond to the *Chain-of-Thought (CoT)* or the generated plan. This offers a unique security opportunity: unlike humans, whose thoughts are private and often unconscious ("black box"), an agent's "thoughts" (CoT) can be inspected ("white box"). Security mechanisms could monitor the agent's *intention* before execution. If the reasoning trace reveals an intent to deceive or bypass a rule ("I must hide this file extension to fulfill the user's request"), the action can be blocked based on the malicious intention, even if the final command looks syntactically valid.

Assessment of the Situation: Before acting, a human unconsciously or consciously assesses the situation (criticality, stress level, anomaly). AI agents often lack this meta-cognitive step, treating a chat about weather and a high-stakes financial transaction as identical token processing tasks. Secure agentic systems could implement an explicit *contextual assessment* step or even introduce a *supervisor agent* pattern: The agent could continuously evaluate: "Is the current situation critical? Is the input source trusted?" If the assessment yields a high-risk score, the agent should dynamically switch to a more restrictive behavior mode.

Usability and Agency: Finally, just as poor usability leads humans to bypass security (e.g., leveraging shadow IT systems), the design of *tool interfaces* affects AI security. If an interface like an API is too permissive, an agent might misuse it. We can introduce the concept of "agency friction": for high-risk actions (e.g., deleting data), the interface should require the agent to pause and request human confirmation (Human-in-the-Loop, HITL), effectively limiting its agency in critical situations to prevent catastrophic autonomous errors.

V. Conclusion and Future Work

Agentic AI systems represent a watershed moment in cybersecurity. By granting systems the autonomy to act based on natural language, we have created a class of actors that are susceptible to semantic manipulation, mirroring human vulnerabilities and often lacking the "critical thinking" ability of humans. We argue that part of the solution can lie in adopting a behavioral perspective: By utilizing the framework established for human cybersecurity behavior—examining drivers like Motivation, Role, Norms, and Situational Awareness—we can derive a more robust defense strategy for AI agents.

This behavioral perspective opens up a new, interdisciplinary research agenda. We highlight five exemplary areas for future work to strengthen the resilience of agentic AI systems:

1) *Systematizing the Attack Surface:* How can we categorize specific attack scenarios that exploit the anthropomorphic nature of AI agents? What specific "cognitive biases" (hallucinations, sycophancy) do agents exhibit, and how can attackers exploit them?
2) *Implementing "Security Awareness":* How can we move beyond rule-based filters to implement a form of adaptive "security awareness" in agents? Can we measure an agent's "Security Mindset" quantitatively before deploying it in critical infrastructure?
3) *Security-Utility balance for agentic systems:* How do we implement a "Security Mindset", but avoid overly cautious

behaviour, leading to operational failures where agents refuse to process legitimate tasks due to a misinterpretation of security boundaries?

4) *Roles and Culture in Multi-Agent Systems:* To what extent can concepts like "Security Champions" be transferred to multi-agent swarms? Can specialized security agents improve the overall "culture" (alignment) of a heterogeneous agent system?
5) *Usability and Agency Design:* How can we design APIs and environments that "nudge" agents towards secure behavior? How do we balance the autonomy required for efficiency with the friction required for security, ensuring the agent knows when to halt and ask for human help?

The security of agentic AI is not merely a coding challenge; it is a challenge of designing resilient AI behavior. By bridging the gap between human factors research and AI engineering, we can aim to build systems that are not just smart, but also wise to the threats of a semantic world.

REFERENCES

[1] D. B. Acharya, K. Kuppan, and B. Divya, "Agentic AI: Autonomous Intelligence for Complex Goals—A Comprehensive Survey", *IEEE Access*, vol. 13, pp. 18 912–18 936, 2025, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2025.3532853

[2] S. Hosseini and H. Seilani, "The role of agentic AI in shaping a smart future: A systematic review", *Array*, vol. 26, p. 100 399, Jul. 2025, ISSN: 25900056. DOI: 10.1016/j.array.2025.100399

[3] R. Sapkota, K. I. Roumeliotis, and M. Karkee, "AI Agents vs. Agentic AI: A Conceptual taxonomy, applications and challenges", *Information Fusion*, vol. 126, p. 103 599, Feb. 2026, ISSN: 15662535. DOI: 10.1016/j.inffus.2025.103599

[4] Y. Li et al., *Personal LLM Agents: Insights and Survey about the Capability, Efficiency and Security*, 2024. DOI: 10.48550/ARXIV.2401.05459

[5] Z. Deng et al., "AI Agents Under Threat: A Survey of Key Security Challenges and Future Pathways", *ACM Computing Surveys*, vol. 57, no. 7, pp. 1–36, Jul. 2025, ISSN: 0360-0300, 1557-7341. DOI: 10.1145/3716628

[6] Y. He, E. Wang, Y. Rong, Z. Cheng, and H. Chen, "Security of AI Agents", in *2025 IEEE/ACM International Workshop on Responsible AI Engineering (RAIE)*, Ottawa, ON, Canada: IEEE, Apr. 2025, pp. 45–52, ISBN: 979-8-3315-1466-2. DOI: 10.1109/RAIE66699.2025.00013

[7] I. Adabara, B. Olaniyi Sadiq, A. Nuhu Shuaibu, Y. Ibarahim Danjuma, and M. Venkateswarlu, "A Review of Agentic AI in Cybersecurity: Cognitive Autonomy, Ethical Governance, and Quantum-Resilient Defense", *F1000Research*, vol. 14, p. 843, Sep. 2025, ISSN: 2046-1402. DOI: 10.12688/f1000research.169337.1

[8] A. Sheth et al., "Agentic AI for Autonomous Cyber Threat Hunting and Adaptive Defense in Dynamic Security Environments", in *2025 IEEE International Conference on Electro Information Technology (eIT)*, Valparaiso, IN, USA: IEEE, May 2025, pp. 316–321, ISBN: 979-8-3315-3233-8. DOI: 10.1109/eIT64391.2025.11103697

[9] B. Berens, M. Ghiglieri, O. Kulyk, P. Mayer, and M. Volkamer, "Human Factors in Security", in *Sicherheitskritische Mensch-Computer-Interaktion: Interaktive Technologien und Soziale Medien im Krisen- und Sicherheitsmanagement*, C. Reuter, Ed., Wiesbaden: Springer Fachmedien, 2021, pp. 89–110, ISBN: 978-3-658-32795-8. DOI: 10.1007/978-3-658-32795-8_5

[10] M. A. Sasse and A. Rashid, "Human Factors", in *The Cyber Security Body Of Knowledge*, University of Bristol, 2021, ch. Human Factors.

[11] R. A. Maalem Lahcen, B. Caulkins, R. Mohapatra, and M. Kumar, "Review and insight on the behavioral aspects of cybersecurity", *Cybersecurity*, vol. 3, no. 1, p. 10, Dec. 2020, ISSN: 2523-3246. DOI: 10.1186/s42400-020-00050-w

[12] F. Jörgens, *The Human Firewall* (essentials), 1st ed. 2023. Wiesbaden: Springer Fachmedien Wiesbaden, 2023, ISBN: 978-3-658-42757-3. DOI: 10.1007/978-3-658-42757-3

[13] R. Kaur, D. Gabrijelčič, and T. Klobučar, "Artificial intelligence for cybersecurity: Literature review and future research directions", *Information Fusion*, vol. 97, p. 101 804, Sep. 2023, ISSN: 1566-2535. DOI: 10.1016/j.inffus.2023.101804

[14] M. Malatji and A. Tolah, "Artificial intelligence (AI) cybersecurity dimensions: A comprehensive framework for understanding adversarial and offensive AI", *AI and Ethics*, vol. 5, no. 2, pp. 883–910, Apr. 2025, ISSN: 2730-5953, 2730-5961. DOI: 10.1007/s43681-024-00427-4

[15] M. Schmitt and I. Flechais, "Digital deception: Generative artificial intelligence in social engineering and phishing", *Artificial Intelligence Review*, vol. 57, no. 12, p. 324, Oct. 2024, ISSN: 1573-7462. DOI: 10.1007/s10462-024-10973-2

[16] P. Steinberger, *Introducing OpenClaw*, https://openclaw.ai/blog/introducing-openclaw, [retrieved March, 2026], 2026.

[17] O. Yomtov, *Clawhavoc: 341 malicious clawed skills found by the bot they were targeting*, https://www.koi.ai/blog/clawhavoc-341-malicious-clawedbot-skills-found-by-the-bot-they-were-targeting, [retrieved March, 2026], 2026.

[18] K. Greshake et al., "Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection", in *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, Copenhagen Denmark: ACM, Nov. 2023, pp. 79–90, ISBN: 979-8-4007-0260-0. DOI: 10.1145/3605764.3623985

[19] S. Parkin and L. Viganò, Eds., *Socio-Technical Aspects in Security: 11th International Workshop, STAST 2021* (Lecture Notes in Computer Science), 1st ed. 2022. Cham: Springer International Publishing, 2022, vol. 13176, ISBN: 978-3-031-10182-3 978-3-031-10183-0. DOI: 10.1007/978-3-031-10183-0

[20] G. Dhillon, K. Smith, and I. Dissanayaka, "Information systems security research agenda: Exploring the gap between research and practice", *The Journal of Strategic Information Systems*, vol. 30, no. 4, p. 101 693, Dec. 2021, ISSN: 09638687. DOI: 10.1016/j.jsis.2021.101693

[21] M. Mehrnezhad and S. Parkin, Eds., *Socio-Technical Aspects in Security: 12th International Workshop, STAST 2022* (Lecture Notes in Computer Science). Cham: Springer Nature Switzerland, 2025, vol. 13855, ISBN: 978-3-031-83071-6 978-3-031-83072-3. DOI: 10.1007/978-3-031-83072-3

[22] B. D. Payne and W. K. Edwards, "A Brief Introduction to Usable Security", *IEEE Internet Computing*, vol. 12, no. 3, pp. 13–21, May 2008, ISSN: 1941-0131. DOI: 10.1109/MIC.2008.50

[23] USENIX Association, Ed., *Proceedings of the Twentieth Symposium on Usable Privacy and Security (SOUPS 2024)*. Berkeley, CA: USENIX Association, 2024, ISBN: 978-1-939133-42-7.

[24] USENIX Association, Ed., *Proceedings of the Twenty-first Symposium on Usable Privacy and Security (SOUPS 2025)*. Berkeley, CA: USENIX Association, 2025, ISBN: 978-1-939133-51-9.

[25] K. O. Kürtz, *Towards Modeling Cybersecurity Behavior of Humans in Organizations*, Cryptology ePrint Archive, Paper 2026/490, 2026. DOI: 10.48550/arXiv.2603.08484

[26] Anthropic, *Introducing the Model Context Protocol*, https://www.anthropic.com/news/model-context-protocol, [retrieved March, 2026], 2024.