# Anthropomorphic Trust - How Artificial Intelligence threatens established information security principles

Niklas Kruse
*University of Applied Sciences Osnabrück*
Germany
E-Mail: niklas.kruse@hs-osnabrueck.de

Celine Richter, Morna Müller, Julian Dreyer
*MUUUH! GmbH*
Germany
E-Mail: {celine.richter, morna.mueller, julian.dreyer}@muuuh.de

***Abstract*—Large Language Models (LLMs) and LLM-based AI agents are increasingly deployed as autonomous actors within corporate value chains, accessing data and tools on behalf of their users. Unlike traditional software assets, they communicate in natural language and are therefore prone to being anthropomorphized. This paper examines, from an interdisciplinary perspective spanning information security, data privacy, and cognitive psychology, how the resulting trust relationship between humans and AI agents undermines established security principles such as need-to-know and least-privilege. We show that users anthropomorphizing these systems tend to grant increasingly broad permissions to them, even though these systems fundamentally lack the capacities that would justify such trust in security-critical domains. This creates novel gaps in confidentiality, integrity, and accountability that current standards like ISO 27001 or the NIST Cyber Security Framework do not sufficiently address. We argue that AI systems should be treated as a distinct third asset category within an Information Security Management System (ISMS) — alongside humans and conventional software — and outline organizational and technical countermeasures, including the controlled de-anthropomorphization of agents.**

***Index Terms*—Artificial Intelligence, LLM, Anthropomorphism, Trust, Information Security, Data Protection**

## I. Introduction

Since their release in 2022, large language models (LLMs) have moved beyond purely academic applications and are now being used throughout the society. More recently, LLMs have been increasingly deployed as agent-based systems that autonomously access data sources, tools, and interfaces [1]. Companies are using such agent based systems to automate processes and gain efficiencies [1, 2]. In this regard, agents become part of a company's Information Security Management System (ISMS) as they interact with company assets. Agents using artificial intelligence (AI) represent a new form of data processing [3]. While they process data as an autonomous or semi-autonomous machine, they interact with the user using natural language. Although the opportunities and ethical risks of AI agents have been extensively discussed in social and academic discourse [1, 3], the implications for information security and corporate governance arising from this new interaction between humans and machines have received little attention to date. Currently research in regard to agent safety is focused on the agents themselves [4, 5] and not on the interaction with other humans as stakeholders in an ISMS. Phenomena such as anthropomorphism and trust in machines present new challenges in this context [6].

To highlight some of the most influential research gaps, this paper discusses the following interdisciplinary effects of current LLM-based AI technologies in a sociodemographic context:

- Information security tradeoffs and risks caused by anthropomorphic AI development
- Organizational risks induced by broad data access granted to AI-chatbots

This paper is structured as follows: Section II provides a broad overview of current related work in the field of anthropomorphism in the context of AI and implicit trust relations being formed between humans and LLMs. Next, section III discusses the social and economic effects of the trust relationships between humans and their chatbot tools. Finally, IV concludes this paper and highlights further future research questions.

## II. Related Work

In respect to this paper being an interdisciplinary study, the related work is based not only on aspects of information security and data privacy, but also on cognitive and psychological interaction between users and AI.

### A. LLMs and AI agents

LLMs are a technology used in natural language processing (NLP). Modern LLMs are based on transformer architectures, which have enabled significant advances in the general field of AI [7]. The transformer architecture incorporates a central mechanism called "self-attention". It is used by the LLMs when breaking down human language into individual tokens and assigning them individual relevance relative to other tokens in the same input sequence. This mechanism enables the AI to identify coherent patterns even in long texts [8]. In a first step, LLMs are pre-trained on large text data sets using a self-supervised objective. The models learn to predict next most probable token given the preceding token context [7]. Through this process, the model can derive syntactic, semantic, and factual information from the training data, which are represented in the internal model parameters. Note that this entire process is built on stochastic principles and probabilities rather than true knowledge. After pre-training, models are typically fine-tuned and aligned to perform in desired behaviors [7]. An

LLM always generates an output based on the input data in the form of a prompt. This generation always occurs token by token. Each output token is drawn from the learned probability distribution, given the prompt and all previously generated tokens [7]. Thus, LLMs represent probabilistic systems whose outputs are not derived from symbolic considerations or an actual knowledge base, but rather from the probability of the tokens in question.

With the omnipresence of AI-chatbots, the more general terms "AI agent" and "agentic-AI" have emerged in academia and the economy alike. AI agents in particular gained significant importance with recent advancements in generative AI [1]. An AI agent is a piece of software that solves a given task with varying degrees of autonomy, accessing additional data sources, contextual information, and tools to interact with its environment, or even other AI agents [2, 1]. While an AI agent is considered a single software object, agentic AI refers to a network of specialized AI agents that interact with one another [2, 3]. LLMs are a key component of an AI Agent as they serve as the reasoning and orchestration component. The model leverages the inference capabilities acquired during pre-training and fine-tuning, interpreting inputs and reasons about tasks. Based on the system prompt of the AI agent, the LLM generates a specialized response to the given task [1]. In contrast to a general LLM only being able to access its trained knowledge [7] an AI agent extends this principle by having access to additional tools, e.g., using Application Programming Interfaces (APIs) [1]. If the LLM determines that further knowledge or interaction beyond its own trained knowledge is necessary, a corresponding call to an available tool is performed. The results of this call serve as additional input for the LLM, which responds accordingly. The collaboration of different AI agents within an agentic AI system significantly reinforces the black-box nature of AI while also providing a more broader application potential. In such systems, AI agents interact with each another thus making reasoning decisions hard to trace [2].

### *B. Anthropomorphism and Human-AI Relationships*

Anthropomorphism refers to the attribution of human characteristics to a non-human entity. This can include human motivations, intentions, and emotions. [9, 10, 11] This effect can be applied to many different objects, which do not necessarily have to be living entities. Anthropomorphism not only leads to inanimate objects being viewed as alive, but also to them being treated with as if they were human [12, 10, 11]. Neuroscience suggests that anthropomorphism draws on the same cognitive processes that support human social cognition [13, 14]. Consequently, anthropomorphism is not an irrational belief, but rather reflects normal human cognitive processes [9]. Fundamentally, anthropomorphism is a process of inductive inference in which perceived characteristics lead individuals to infer similarities between themselves and another entity, thereby engaging in analogical reasoning [15]. The SEEK model provides a clearer understanding of the conditions that trigger anthropomorphism [9]. To interpret the events around them, humans rely on their own worldview. When confronted with unfamiliar influences, the human brain projects homocentric and egocentric human behavioral models onto those influences. The greater the similarity between these influencing factors and the behavioral models, the more likely anthropomorphism is to occur [9]. One of the most significant factors of human interaction is our dependence on language functioning as an exchange medium for instructions, emotions, and knowledge [16].

As discussed in the previous section, AI-chatbots are intentionally trained on generally available, human-written texts and inputs, thus resulting in an output that also sounds human at first glance. AI systems are therefore particularly prone to being anthropomorphized by humans [17], as they provide strong human-like communication patterns. Despite the probabilistic generation of additional tokens during output, even experts find it difficult to predict the behavior of an LLM without a deep look into its parameters. In addition to the widespread use of AI mentioned in section I, this further increases the impression of an AI black-box system when interacting with natural language.

The degree of anthropomorphism can also depend on a number of other factors stemming from both the AI user and the AI's design [17]. Trust in AI systems can be divided into two categories: trust in the user's and the system's technical functionality, and human-like trust in the AI system itself [18, 19]. Due to its mimicry of human-like mental states anthropomorphism leads people to develop trust in the AI system itself [20, 21, 22, 23], even though, due to its probabilistic nature and technical limitations, the system is not capable of acting in a trustworthy, reasonable manner on its own. This is evident in the way relationships between people and chatbots develop. When using chatbots, a dynamic relationship develops between the user and the AI system [24, 25], which evolves as the chatbot is used [26]. Although the closeness of the relationship also depends on the frequency and intensity of use [24]. Some users anthropomorphize the AI system even before using it for the first time [25], calling it nicknames, consulting its help in typical human-focused interactions, or creating an empathic bond between the AI system and the user.

Trust, as an integral part of the relationship, towards the AI is built among users through their interaction with the system. However, as Skjuve et al. point out, if the system becomes unstable and behaves unexpectedly even for a brief moment the established users' trust in the AI system can deteriorate directly [25]. In such cases, the technical malfunction disrupts the anthropomorphic effect. These malfunctions can undermine users' trust in a chatbot to such an extent that they are subsequently less willing to provide the same level of data, i.e., private or even secret information, to the AI system [25]. Attributing human-like characteristics to AI systems increases users' trust, which is is further reinforced through repeated interaction. As such anthropomorphized AI systems are not merely perceived by humans as technical tools. In such cases, users view the AI system as a genuine assistant.

### C. Information Security and Access Management

The rise of AI in general economy led to new pathways and operational improvements both for attackers and defenders of information security. In general, IT-Security aims to protect assets such as business information, hardware-assets, or intellectual property in three basic dimensions: Confidentiality, integrity, and availability (CIA-Principle). Furthermore, these principles can be extended to include the accountability and non-repudiation of information, depending on the concrete asset. Information security aims to secure these assets by providing clear business policies, technological measures, and monitoring strategies.

Two of the basic concepts of information security are the principles of "Need-to-Know" and "Least-Privilege", both originating from the domain of Identity and Access Management (IAM) [27]. Assigning privileges in cyber-physical system shall only be conducted on a need to know basis such that users can only access information that they need for their intended use case. In addition, users must only have the least possible privileges to perform a task successfully. For example, if a user wants to read a text document in a shared directory, granting general administrative access to the directory violates both the need-to-know and least privilege principles. By granting only read-permissions to the intended file, both concepts are fulfilled.

The general aim of both concepts is to ensure that no malicious actor or even the user can exploit the extended privileges for harm. In an event where an information security breach occurs, access to the affected assets must be traced back to individual users accessing the assets with their respective access rights. This creates a fully accountable environment where security gaps can be identified with great details but also increases the overall requirement for planning and operating a fine-granular access structure.

## III. Risk of Anthropomorphic trust in IT Security

Many businesses use agentic AI systems in their value chain, e.g., in software development processes, consulting operations, or financial analytics [3, 28]. In most scenarios, the users of these systems use the LLMs to generate new inspirations and ideas for their work but also generate entire documents, scripts, or media that are shipped to customers as they are, even without a posterior validation of the AI-generated output in these files. The chatbot is considered to be a valuable sparring partner, taking away major workload from the user and reducing stress at work just like a junior-colleague would do.

Therefore, users establish a trusted relationship with their AI-agents and rely on their outputs to be valid and integer. Studies show, that over a prolonged interaction with the AI systems, users establish an increasingly stronger reliance on the output and also assign a higher informational confidence in the validity of the AI-generated output [26, 25, 24]. This is a particularly paradoxical effect since the degree of AI-output validity does not change whereas the felt-degree of validity by the user increases over time.

This behavior is intrinsic to humans since we expect a co-worker to learn and progress over time and apply the newly learned facts and skills in new scenarios and also learn from errors [29]. While this is true for humans in most cases, LLM-based AI-systems do not behave in the same way. They do not learn from errors and do not improve their output validity over time just by mere interaction, thus creating a discrepancy in the human expectation and the experienced reality also due to the inherent misunderstanding of the technical reality of AI.

This is accompanied by the effect of assigned output validity just by evaluating the eloquence of the AI-generated output. Since the underlying models are trained to produce tokens that sound familiar and professional, humans tend to assign well-sounding phrases a higher informational correctness, even if the generated output is bogus. [30]

### A. AI agent permissions

Both aforementioned cognitive effects lead to a fundamental problem in information security: AI-agents have capabilities to interact with other tools via interfaces that are administered mostly by humans who are susceptible to these effects. When interacting with specialized AI-agents for the first time, human users may choose to grant the AI-agent access to only a defined set of files, thus complying to the need-to-know principle. Over time, when the interaction with the AI-agent increases and the user has created a stronger trust relationship to the human-sounding AI-agent, the user is more willing to grant the AI-agent broader access right, i.e., to administer files without supervision. Note, that this behavior is analogous to supervising a new colleague in his first few months at a new job. The user assigns a higher confidence to the AI output than at the beginning. Understanding the way AI agents work through a human lens can obscure the fact that they lack the fundamental capacities required to justify such trust, such as reliability, predictability, and accountability, creating a critical security gap.

Since LLM based AI-agents operate entirely on probabilistic distributions, granting global administration privileges have a non-negligible chance to be exploited for malicious tasks at some point in time. With this, accountability for leaked credentials, deleted databases, or integrity breaches cannot be attributed to a human person anymore but rather to an autonomous AI-system without a conscience.

Slow but steady, potentially unaware anthropomorphism of AI-agents can cause users to grant unnecessarily broad privileges to these autonomous systems, thus jeopardizing overall information security. Users cannot rely on the informational integrity and validity of AI-output, particularly in highly sensitive work scenarios where confidential or critical assets are involved.

Traditional corporate governance approaches have typically treated human actors, infrastructure, and software assets as separate entities. Management systems must be adapted to respond to this new type of relationship between people and software. Anthropomorphism and the resulting evolution of the

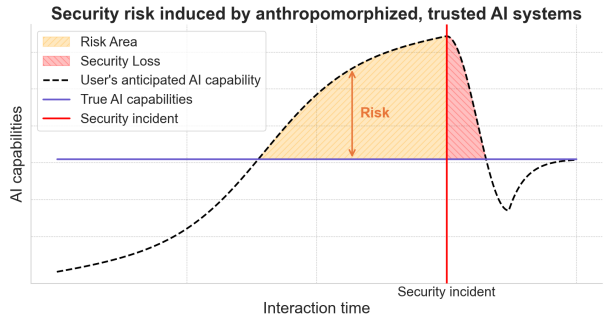


Fig. 1: Induced security risks due to faulty AI-capability assumption by the user interacting with the AI system

relationship mean that, within a company's information system, agents inherit human-like vulnerabilities while retaining operational strengths of AI. They represent virtual colleagues who potentially have been granted too many or overly broad permissions and access to data. An ISMS has to treat these entities particularly different than traditional software systems, as their way of interaction forms an anthropomorphism with the human user as they seem trustworthy. Hence, these AI systems need to undergo a broad regulation in terms of IAM, logging and monitoring, and penetration testing.

### *B. Consequences for ISMS*

Traditional ISMS guidelines such as the ISO27001 or NIST Cyber Security Framework provide dedicated regulations and strategies to secure both human interactions with information systems and software development processes. However, only in the field of human security, awareness of the workers for compliance is considered whereas software products do not follow the same principles. With AI, both domains need to be merged since AI agents undergo an anthropomorphism, as discussed earlier.

Therefore, we propose to apply a different approach towards securing human AI interactions in the information security and data privacy domain. Traditional management systems must recognize that agents are just as opaque and highly variable actors within an ISMS as human employees can be, other than traditional software systems. In corporate governance, humans play an integral role in the interaction between information assets and their accompanying processes. Humans, however, are volatile and prone to making errors, which management systems do consider and cope for. Strategies for business continuity management and recovery plans are formed, making a security incident caused by a human less impactful.

With the introduction of AI, these strategies will not be applicable in that way, since AI systems lack two central human features: Conscience and morale. While humans overthink potentially harmful operational decisions, AI agents will execute them without any awareness for consequences or impact. Hence, ISMS strategies can not be based on the approach of completely eliminating errors and reducing risk, but should assume that they will occur, just as is the case with human errors.

Thus, while ISMS include strategies and policies for both human- and software-security, we propose to include a distinctly different third category with AI-security. Treating AI as a human assets leaves a security gap in software development details while treating AI as a normal software asset neglects the anthropomorphic effect during human interaction. Hence, developing a policy for AI governance and security is required.

Organizational governance should assume that users will anthropomorphize the agents, as some users already do from the initial contact onward [25]. Figure 1 highlights this effect by indicating the increased risk for the occurrence of a security event with rising trust in an anthropomorphized AI agent. An ISMS should also implement organizational mechanisms that actively counteract this effect. One solution could be to induce targeted and controlled malfunctions in the agents, rendering them less human-like. This effect is also shown in fig. 1 after a security incident occurs, either orchestrated using a designed scenario, or by a real attack. In that case, real corporate loss sets in as well as an implicit correction of the true AI's capabilities assumed by the user [26, 25, 24].

## IV. Conclusion and Future Work

The anthropomorphization of AI agents, driven by human behavior, creates a relationship between AI and users that poses entirely new challenges for data protection and information security. This applies not only to technical challenges, but in the way we view these systems as stakeholders in a corporate governance system. This has received little attention in academic discourse to date. The discussion in this paper focuses primarily on the security implications of this particular effect.

The results show that even established ISMS standards lack formal strategies to counteract the potentially harmful aspects of the implicit anthropomorphism induced by human AI interactions. By applying a distinct organizational control or AI strategy to these kinds of interactions, corporate environments can integrate AI information security in the established ISMS processes while treating AI systems as both human- and software-like assets.

Nevertheless, from this description of the connection between anthropomorphism and the resulting relationship with the user, further research questions and open issues can be derived:

- How can operational information security tradeoffs and risks, caused by anthropomorphic AI development, be mitigated effectively?
- Who is legally responsible for violations given the unique relationship between humans and AI?
- Does the European Union's AI-Act effectively cope with anthropomorphism in AI systems?

In light of the described effects of anthropomorphism, this raises the question of whether, in such human-AI interactions, humans can make objective decisions at all, or whether this objective control function is lost over time as a relationship develops.

## References


[1] R. Sapkota, K. I. Roumeliotis, and M. Karkee, "Ai agents vs. agentic ai: A conceptual taxonomy, applications and challenges," *Information Fusion*, p. 103599, 2025.

[2] A. K. Pati, "Agentic ai: A comprehensive survey of technologies, applications, and societal implications," *IEEE Access*, vol. 13, pp. 151 824–151 837, 2025.

[3] S. Hosseini and H. Seilani, "The role of agentic ai in shaping a smart future: A systematic review," *Array*, vol. 26, p. 100399, 2025.

[4] A. Chhabra, S. Datta, S. K. Nahin, and P. Mohapatra, "Agentic ai security: Threats, defenses, evaluation, and open challenges," *arXiv preprint arXiv:2510.23883*, 2025.

[5] S. Raza, R. Sapkota, M. Karkee, and C. Emmanouilidis, "Trism for agentic ai: A review of trust, risk, and security management in llm-based agentic multi-agent systems," *AI Open*, 2026.

[6] G. Abercrombie, A. C. Curry, T. Dinkar, V. Rieser, and Z. Talat, "Mirages. on anthropomorphism in dialogue systems," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 4776–4790.

[7] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, vol. 1, no. 2, pp. 1–124, 2023.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[9] N. Epley, A. Waytz, and J. T. Cacioppo, "On seeing human: a three-factor theory of anthropomorphism." *Psychological review*, vol. 114, no. 4, p. 864, 2007.

[10] G. Airenti, "The cognitive bases of anthropomorphism: from relatedness to empathy," *International Journal of Social Robotics*, vol. 7, no. 1, pp. 117–127, 2015.

[11] A. Salles, K. Evers, and M. Farisco, "Anthropomorphism in ai," *AJOB neuroscience*, vol. 11, no. 2, pp. 88–95, 2020.

[12] N. Epley, A. Waytz, S. Akalis, and J. T. Cacioppo, "When we need a human: Motivational determinants of anthropomorphism," *Social cognition*, vol. 26, no. 2, pp. 143–155, 2008.

[13] F. Castelli, F. Happé, U. Frith, and C. Frith, "Movement and mind: a functional imaging study of perception and interpretation of complex intentional movement patterns," in *Social neuroscience*. Psychology Press, 2013, pp. 155–169.

[14] H. Cullen, R. Kanai, B. Bahrami, and G. Rees, "Individual differences in anthropomorphic attributions and human brain structure," *Social cognitive and affective neuroscience*, vol. 9, no. 9, pp. 1276–1280, 2014.

[15] P. Bartha, *Analogy and Analogical Reasoning*, Fall 2024 ed. Metaphysics Research Lab, Stanford University, 2024.

[16] S. Rabiah, "Language as a tool for communication and cultural reality discloser," 2018.

[17] R. Miranda, "Seemingly intelligent: The inherent dangers of mistaking a for i. a study on the effects of language choice and anthropomorphism in ai." 04 2024.

[18] H. Choung, P. David, and A. Ross, "Trust in ai and its role in the acceptance of ai technologies," *International Journal of Human–Computer Interaction*, vol. 39, no. 9, pp. 1727–1739, 2023.

[19] J. A. Everett, S. Claessens, T.-D. Knöchel, and M. G. Reinecke, "Principles for understanding trust in artificial intelligence," *Nature Reviews Psychology*, pp. 1–14, 2026.

[20] E. J. De Visser, S. S. Monfort, R. McKendrick, M. A. Smith, P. E. McKnight, F. Krueger, and R. Parasuraman, "Almost human: Anthropomorphism increases trust resilience in cognitive agents." *Journal of Experimental Psychology: Applied*, vol. 22, no. 3, p. 331, 2016.

[21] T. Araujo, "Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions," *Computers in human behavior*, vol. 85, pp. 183–189, 2018.

[22] V. Herdel and J. R. Cauchard, "Anthropomorphism as a dimension of trust in conversational agents," in *CHI '24: Building Trust in CUIs – From Design to Deployment*, 2024, pp. 1–6.

[23] P. Ribino, "The role of politeness in human–machine interactions: A systematic literature review and future perspectives." *Artificial Intelligence Review*, vol. 56, 2023.

[24] M. Chandra, J. Hernandez, G. Ramos, M. Ershadi, A. Bhattacharjee, J. Amores, E. Okoli, A. Paradiso, S. Warreth, and J. Suh, "Longitudinal study on social and emotional use of ai conversational agent," *arXiv preprint arXiv:2504.14112*, 2025.

[25] M. Skjuve, A. Følstad, K. I. Fostervold, and P. B. Brandtzaeg, "A longitudinal study of human–chatbot relationships," *International Journal of Human-Computer Studies*, vol. 168, p. 102903, 2022.

[26] ——, "My chatbot companion-a study of human-chatbot relationships," *International Journal of Human-Computer Studies*, vol. 149, p. 102601, 2021.

[27] F. B. Schneider, *Least Privilege and More*. New York, NY: Springer New York, 2004, pp. 253–258. [Online]. Available: https://doi.org/10.1007/0-387-21821-1_38

[28] D. A. S. George, "The agent economy: How autonomous ai systems are restructuring knowledge work, capital allocation, and the architecture of enterprise value," *Partners Universal Innovative Research Publication*, vol. 4, no. 1, p. 1–41, Mar. 2026. [Online]. Available: https://www.puirp.com/index.php/research/article/view/153

[29] G. Alessandri, L. Borgogni, and D. M. Truxillo, "Tracking job performance trajectories over time:

A six-year longitudinal study," *European Journal of Work and Organizational Psychology*, vol. 24, no. 4, pp. 560–577, 2015. [Online]. Available: https://doi.org/10.1080/1359432X.2014.949679

[30] T. Kleinert, M. Waldschütz, J. Blau, M. Heinrichs, and B. Schiller, "Ai outperforms humans in establishing interpersonal closeness in emotionally engaging interactions, but only when labelled as human," *Communications Psychology*, vol. 4, no. 1, p. 23, 2026. [Online]. Available: https://doi.org/10.1038/s44271-025-00391-7