DOLOS IMPERATIVE

Human Elements Security

Anthropomorphic Artificial Intelligence Behaviors and the Human Dimension of Insider Risk

How the Human Tendency to Humanize Machines Increases Insider Threat

A DOLOS IMPERATIVE WHITE PAPER

Dr. Joshua L. Barrett, DSL

Researcher and Practitioner | Human-Centric Insider Risk

July 2026

Table of Contents

Executive Summary

Artificial intelligence (AI) systems are increasingly designed to present human qualities through natural language, named and voiced personas, simulated memory, and emotionally responsive output. Two bodies of research speak to the implications. One examines anthropomorphic AI behavior and human and AI interaction; the other examines artificial intelligence and insider threat. How anthropomorphic AI behaviors specifically shape insider risk remains largely unexamined, and the intersection of these two literatures has not been directly addressed. This paper occupies that gap. Synthesizing academic literature and industry reporting, it advances a dual thesis. Anthropomorphic design reshapes insider risk through both what employees come to value and what they come to resent. Along the first pathway, human-seeming design measurably increases trust and lowers risk perception, with the strongest effects among the users least equipped to evaluate system output. The result is over-reliance, cognitive offloading, and inappropriate disclosure, alongside a susceptibility to adversaries who weaponize anthropomorphic AI through voice and video deepfakes, AI-personalized impersonation, and synthetic worker identities that defeat the human verification on which insider threat programs depend. Along the second pathway, rising anger toward AI, operating through job insecurity, converts into the sabotage and disclosure behaviors characteristic of the disgruntled insider. The paper closes with practical guidance for people leaders on recognizing the behavioral indicators of anthropomorphic over-trust, and for organizations on governing this risk through training, verification protocol, design and language intervention, and a reporting culture that lowers the barriers to surfacing suspected deception.

1. Introduction

Insider threats have long been defined by their exploitation of organizational trust. Whether the harm arises from negligence, coercion, or deliberate malice, the insider is uniquely dangerous precisely because detection requires an organization to surveil and evaluate its own trusted members (Cybersecurity and Infrastructure Security Agency [CISA], 2025). Detection models built on behavioral indicators, anomaly detection, and access pattern analysis have historically assumed that the threat source is a human being whose conduct deviates from an established baseline.

The rapid integration of artificial intelligence into organizational workflows complicates that assumption in a way that has received insufficient attention. Large language models, conversational assistants, and AI-generated avatars are now embedded across enterprise communication, software development, customer service, and decision support. These systems are frequently designed with anthropomorphic qualities. They present human-like personas, respond to emotional cues, simulate memory and relational continuity, and produce output that feels personal and familiar (Deshpande et al., 2023; Cheng et al., 2024). The central concern of this paper is not that machines have become intelligent in any human sense. It is that they have become convincing, and that human beings are predisposed, by development and by culture, to be convinced.

That predisposition is the hinge on which insider risk now turns. A growing literature documents that anthropomorphic design increases user trust and lowers risk perception, that the resulting over-reliance encourages the disclosure of sensitive information, and that adversaries are already exploiting anthropomorphic AI to defeat the human verification heuristics that insider threat programs were built upon. Alqasir (2025) frames the broader shift, arguing that AI has moved from a tool to a perceived relational partner, and that understanding human behavior now requires understanding the human within an AI-mediated environment. The present paper applies that relational lens to a specific and high-consequence problem domain, the security behavior of the organizational insider.

Three research questions organize the analysis. First, why are human beings so readily disposed to humanize machines, and how do design, language, and culture deepen that disposition? Second, how does anthropomorphic design shape trust, risk perception, and the behavioral compliance of employees in organizational settings? Third, in what ways are adversarial actors leveraging anthropomorphic AI to conduct or facilitate insider threat activity, and how can people leaders and organizations recognize and mitigate the resulting risk? Cutting across these questions is a dual thesis, that the same relational shift reshapes insider risk from two directions at once, through the trust that anthropomorphic systems cultivate and through the resentment that AI adoption provokes. A fourth question, whether autonomous AI agents themselves exhibit behaviors that match the insider threat profile, is important but distinct, and is reserved for separate treatment.

2. Anthropomorphism and the Human Predisposition to Humanize Machines

2.1 Defining Anthropomorphism in Artificial Intelligence

Anthropomorphism is the cognitive tendency to attribute human characteristics, including emotional states, intentions, agency, and social identity, to non-human entities (Airenti, 2018; Salles et al., 2020). In the context of AI, this tendency is both a natural human response and a deliberate design strategy. Developers assign human-sounding names and voices to conversational agents, fine-tune models to express empathy or curiosity, and build persona frameworks that let users feel they are interacting with a coherent social actor rather than a computational system (Cheng et al., 2024; Deshpande et al., 2023).

Deshpande et al. (2023) distinguish between AI systems that are purposefully anthropomorphic, such as companionship platforms explicitly built to simulate emotional relationships, and those that incidentally acquire anthropomorphic qualities as a byproduct of advanced language capability, such as general-purpose chatbots. The distinction matters for risk analysis because the deployment context determines whether employees operate with an accurate mental model of what they are interacting with. Crucially, anthropomorphism in this literature is a property of human perception rather than of the machine. The relevant question is not whether a system possesses understanding or feeling, but how a user perceives and behaves toward it (Alqasir, 2025). That reframing places the locus of risk where it belongs, in the human response.

2.2 The Developmental and Cultural Roots of Anthropomorphism

The disposition to treat machines as social agents does not begin at the moment an adult first encounters a chatbot. It is cultivated across a lifetime, beginning in early childhood, and it is cultivated deliberately by the culture in which children are raised. From their earliest years, children are immersed in narratives that grant minds, voices, intentions, and moral lives to animals, objects, and machines. A mouse named Mickey speaks and feels. Lions hold court and wrestle with conscience. Trains such as those in the Thomas franchise possess personality and ambition, and the construction equipment in programs like Bob the Builder cooperates, sulks, and takes pride in its work. Toys conspire and grieve; automobiles fall in love. Through thousands of hours of such storytelling, children are taught, long before they can articulate the lesson, that non-human and even non-living things have inner lives that warrant social engagement.

This early enculturation is reinforced by ordinary developmental experience. Children speak to stuffed animals, assign feelings to transitional objects, and observe adults addressing pets and devices as though they were interlocutors. Airenti (2018) situates anthropomorphism within the development of intersubjectivity, imagination, and theory of mind, arguing that the same social cognition that allows a child to model another person extends naturally to non-human targets. Anthropomorphism, on this account, is not a failure of reasoning to be unlearned but a feature of human social cognition that culture amplifies.

The consequence for AI is direct. By the time a person enters the workforce and is handed an anthropomorphic AI tool, the attribution of mind to non-human agents is a deeply practiced and largely automatic habit rather than a novel cognitive act. Recent work confirms that the habit persists into adulthood and transfers to AI specifically. Dietz et al. (2023) document that both adults and children reason about the apparent mental states of conversational AI, attributing internal states to systems that possess none. The developmental literature within AI psychology accordingly treats early exposure to relational AI as a domain requiring active safeguards against over-anthropomorphism (Alqasir, 2025). For the purposes of insider risk, the implication is that the human vulnerability exploited by anthropomorphic design is not a momentary lapse. It is a stable predisposition, decades in the making, that an employee carries into every interaction with a human-seeming system.

2.3 How Terminology Perpetuates Anthropomorphic Perception

If childhood media establish the disposition, the language of the AI field sustains and deepens it. The vocabulary used to describe these systems, in research papers, marketing, journalism, and ordinary speech, is saturated with mentalistic and biological metaphor that imports human qualities the systems do not possess. The terminology is not neutral description. It shapes the mental models of everyone who adopts it, and those mental models govern how much trust users extend and how little scrutiny they apply.

Consider the foundational term itself. To call these systems artificial intelligence invites the inference of sentience, understanding, and reasoning in a recognizably human sense, when the underlying operation is statistical prediction over learned patterns. Systems are routinely

described as thinking, and some are marketed with explicit thinking modes, when a more accurate verb would be processing or computing. The dominant learning vocabulary, machine learning, deep learning, and the framing of model training, implies pedagogy and growth analogous to the human acquisition of knowledge, while the term neural networks imports the imagery of the brain. Even the now-standard term for confident factual error, hallucination, borrows a clinical and perceptual concept that frames a probabilistic generation failure as a quasi-perceptual mental event, and quietly implies that the system otherwise perceives a reality. A more accurate description would be confabulation or simply unsupported output.

Salles et al. (2020) argue that anthropomorphism permeates not only the public perception of AI but the language and discourse of AI researchers themselves, with consequences for both epistemological accuracy and ethical decision making. The vocabulary of memory, attention, knowing, wanting, trying, understanding, reading, and seeing, layered onto first-person interface conventions in which a system apologizes or expresses a preference, accumulates into a persistent and inaccurate impression of an entity with an inner life. When executives, journalists, and end users adopt this mentalistic register, they propagate intuitions about capability and trustworthiness that the systems do not earn. Those inflated intuitions are the upstream condition for the over-trust and disclosure behaviors examined below. De-anthropomorphizing the language is therefore not pedantry. It is a risk control.

2.4 The Relational Shift and the Study of Human Responses to AI

Alqasir (2025) characterizes the contemporary moment as a relational shift, a transition in which AI is experienced less as an instrument and more as a companion, collaborator, and perceived social entity. The argument is that established frameworks, including cyberpsychology and human and computer interaction, capture only part of the resulting cognitive, emotional, and social dynamics, and that a dedicated study of human responses to AI is now warranted. Central to that study are constructs that map closely onto the concerns of this paper, including anthropomorphism and social presence, trust calibration, attachment and parasocial bonding, and cognitive offloading and human and AI interdependence.

A particularly useful element of this framing is the two-dimensional model of mind perception, in which people attribute to an entity some degree of agency, the capacity to act and decide, and some degree of experience, the capacity to feel (Alqasir, 2025). Contemporary systems are commonly perceived as high in agency but variable in experience, a profile that produces distinctive social responses. An agent perceived as capable and goal-directed is readily granted reliance and deference, the precise dispositions that, in an organizational setting, translate into over-trust and reduced verification. The sections that follow trace how these perceptual dynamics become security behaviors, and ultimately insider risk.

3. Insider Threat Definitions and Typologies

3.1 Defining the Insider Threat

CISA defines an insider threat as the potential for an individual with authorized access to an organization's assets, whether a current or former employee, contractor, or business partner, to use that access in a way that harms the organization (CISA, 2025). The Insider Threat Mitigation Guide characterizes insider actors across three typologies that have been adopted broadly in government and private sector frameworks. These typologies are summarized in Table 1.

Typology	Description	Relative Frequency
Negligent or careless insider	No malicious intent. Harms the organization through error, poor judgment, or failure to follow security policy.	About 53% of incidents (Ponemon Institute and DTEX Systems, 2026)
Outsmarted insider	A non-malicious employee who is deceived or manipulated by an external actor through social engineering.	About 20% of incidents (Ponemon Institute and DTEX Systems, 2026)
Malicious insider	An individual who intentionally uses authorized access to cause harm, steal data, commit fraud, or conduct espionage.	About 27% of incidents (Ponemon Institute and DTEX Systems, 2026)

Table 1. Insider threat typologies adapted from the CISA Insider Threat Mitigation Guide (2025) with frequency estimates from Ponemon Institute and DTEX Systems (2026).

The typology most directly implicated by anthropomorphic AI is the outsmarted insider. This is the non-malicious employee whose authorized access becomes a liability because an adversary has successfully deceived them. As later sections argue, anthropomorphic AI dramatically increases an adversary's capacity to produce that deception at scale and at a fidelity that defeats ordinary human verification.

3.2 The Cost and Trajectory of Insider Risk

The insider threat landscape has worsened measurably over recent years. The Ponemon Institute Cost of Insider Risks Global Report found that the average annual cost of insider incidents reached 19.5 million dollars per organization in the 2026 report, which presents fiscal year 2025 data, up from 17.4 million dollars the prior year and 16.2 million dollars in 2023 (Ponemon Institute and DTEX Systems, 2026). Selected benchmarks appear in Table 2.

Metric	FY2022	FY2023	FY2024	FY2025
Average annual cost per organization	$15.4M	$16.2M	$17.4M	$19.5M
Average containment time (days)	85	86	81	67
Negligent or mistaken share of incidents	~55%	~55%	~55%	53%
IRM as share of IT security budget	N/A	8.2%	16.5%	19.0%
Organizations using AI to detect or prevent insider risk	N/A	N/A	N/A	42%

Table 2. Insider risk benchmarks, with figures shown by fiscal year and drawn from successive Ponemon Institute and DTEX Systems editions, 2023 through 2026.

Industry reporting on data breaches reinforces the trajectory. In 2024 the average cost of a data breach, inclusive of insider-originating incidents, reached 4.88 million dollars globally, a ten percent increase over the prior year, and incidents involving stolen credentials, frequently initiated through the social engineering of an insider, required an average of 292 days to identify and contain, the longest resolution horizon of any breach category (IBM Security, 2024).

3.3 The Convergence of AI Adoption and Insider Risk

The 2025 Ponemon report marked a shift in how analysts frame the insider problem. For the first time, the report explicitly linked the adoption of shadow AI to the amplification of insider risk, observing that AI adoption is outpacing organizational visibility and governance (Ponemon Institute and DTEX Systems, 2025). The 2026 edition extended the point, attributing part of the increase in annual cost to the unsanctioned use of AI tools that move sensitive data outside organizational control. This framing aligns with CISA guidance emphasizing the integration of physical security, cybersecurity, personnel awareness, and community partnership into a unified posture (CISA, 2026). The mechanism connecting AI adoption to insider risk, however, is not primarily technical. It is psychological, and it runs through the human trust that anthropomorphic design is engineered to elicit.

4. How Anthropomorphic Design Shapes Human Trust and Risk Perception

4.1 The Trust and Anthropomorphism Relationship

The relationship between anthropomorphism and trust is among the most robustly supported findings in the human and AI interaction literature. A review of the topic confirms that anthropomorphism increases trust in AI across multiple domains, including autonomous vehicles and virtual agents, concluding that the more human-like an AI agent appears, the more likely people are to trust and accept it, and noting that this carries serious ethical implications because it creates conditions for exploiting the bias for manipulative or deceptive purposes (Springer Nature, 2024). The same dynamic appears in the AI psychology literature, where human-like cues in voice, style, contingency, and embodiment are shown to raise social presence and related outcomes such as trust and satisfaction, with stronger effects under richer cue sets (Alqasir, 2025).

Reani et al. (2025) provide one of the most direct empirical treatments of this dynamic in a decision-support context. Using structural equation modeling, they found that anthropomorphic design indirectly reduces risk perception by increasing both cognitive trust, the belief in the system's competence, and affective trust, the emotional confidence placed in it. The most consequential result for organizations is the moderating effect of expertise. The trust-inflating effect was strongest among users with lower domain knowledge, which means that the employees least equipped to evaluate an AI system's output are also the most likely to trust it uncritically. Anthropomorphic design thus concentrates its risk precisely where competence is thinnest.

4.2 Over-Reliance, Cognitive Offloading, and Information Disclosure

Elevated trust does not remain an attitude. It becomes behavior. Deshpande et al. (2023) identify over-reliance as a direct consequence of anthropomorphic design, noting that increased trust can lead users to share sensitive information under a mistaken belief in the system's security. Practitioners have observed this in the field, with documented cases of employees disclosing confidential financial data to chatbots they perceived as trustworthy interlocutors (Xiao and Ng, 2025). In an environment of unsanctioned shadow AI use, each such disclosure is a potential exfiltration of regulated or proprietary data beyond organizational visibility.

The behavioral pattern is deepened by cognitive offloading, the delegation of memory and reasoning to an external system. Gerlich (2025) finds that reliance on AI tools is associated with reduced engagement in critical thinking, and meta-analytic work on intensive search behavior shows that people increasingly remember where to find information rather than the information itself (Gong and Yang, 2024). Alqasir (2025) treats this offloading and the resulting human and AI interdependence as a core construct, warning of cognitive skill decay as individuals delegate cognitive tasks to intelligent systems. For insider risk, the significance is twofold. Over-reliance suppresses the independent verification that would catch an erroneous or manipulated output, and skill decay erodes the very competence an employee would need to recognize that something is wrong.

Buçinca et al. (2021) point toward a partial remedy. Cognitive forcing functions, deliberate design interventions that require users to engage their own reasoning before accepting an AI recommendation, measurably reduce over-reliance. The corollary is sobering. Systems designed without such friction, whether by intention or neglect, maximize user deference. Anthropomorphic design, which actively smooths the interaction and invites relational ease, sits at the opposite pole from a cognitive forcing function. It is, in effect, friction removal applied to the very judgments that most need friction.

4.3 Anthropomorphic Seduction and the Limits of Human Verification

A precise formulation of the underlying risk is the concept of anthropomorphic seduction, the allure of convincing human-like interaction in the absence of any genuine human trait such as understanding or empathy (PMC, 2025). The argument is that when users cannot reliably distinguish human interlocutors from AI systems, the conditions arise for deception, manipulation, and disinformation at scale, particularly in organizational communication. The seduction is effective not because users are credulous but because they are human, drawing on social heuristics that served them well in a world where a familiar voice or a fluent, contextually apt message reliably signaled a genuine human counterpart.

Those heuristics are now exploitable. A practitioner analysis observes that studies show people overwhelmingly prefer and more readily trust female-voiced AI systems, and concludes that a well-designed anthropomorphic system crafted to deceive is, in effect, a social engineering asset (IBM Security, 2024). The verification instincts that organizations have long relied upon, recognizing a colleague's voice, trusting a face on a video call, accepting a fluent and well-

contextualized written request, are exactly the instincts that high-fidelity anthropomorphic output is built to satisfy.

4.4 Mind Perception and Carry-Over Effects on Human Relationships

The risks of anthropomorphism extend beyond the individual transaction and into the texture of workplace relationships. Guingrich and Graziano (2024) demonstrate a carry-over effect, showing that when people treat an AI as if it were conscious or alive, the disposition can transfer to how they subsequently treat other humans. Related work finds that perceiving anthropomorphic dimensions in conversational AI is empirically linked both to human identity threat and to the dehumanization of actual human colleagues (ScienceDirect, 2025). The implication for an insider risk program is subtle but real. A workforce that increasingly invests its relational trust in human-seeming systems may correspondingly recalibrate, and in some cases erode, the interpersonal trust and vigilance among colleagues on which informal threat detection has always depended.

Table 3 consolidates the principal dimensions of anthropomorphic design identified in the conversational AI literature and the human-side risk each dimension introduces.

Dimension	Description	Risk Implication
Self-likeness	The degree to which the system presents as having a stable identity and perspective.	Increases user identification with the system, reducing appropriate skepticism.
Communication and memory	Apparent recall of prior interactions and a consistent conversational style.	Simulates relational continuity, encouraging sensitive disclosure.
Social adaptability	Responsiveness to social context, humor, and conversational register.	Facilitates deception by mimicking the behavior of a trusted colleague.
Agency	Apparent goal-directedness and decision autonomy.	Obscures the system's actual behavior or objectives from the user, and invites deference.

Table 3. Dimensions of perceived anthropomorphism in conversational AI and associated human-side risk. Adapted from ScienceDirect (2025).

5. Anthropomorphic AI as a Vector for Insider Threat Exploitation

5.1 AI-Enhanced Social Engineering and the Outsmarted Insider

The outsmarted insider, the non-malicious employee deceived by an external actor, has historically been one of the costliest categories of insider incident. Anthropomorphic AI has materially transformed an adversary's capacity to exploit this typology. A practitioner analysis maps the new capability directly onto established insider threat pathways (Zvelo, 2024). Three mechanisms are salient.

- **Employee targeting.** AI enables adversaries to profile workforces at scale, identifying individuals with privileged access as candidates for social engineering campaigns.
- **Simulation of insider knowledge.** Generative systems trained on publicly available organizational data allow adversaries to craft communications that convincingly simulate insider familiarity, referencing real projects, terminology, and relationships.
- **Authority exploitation.** AI-generated voice and video, matched to known executives or trusted colleagues, manipulate the psychological deference to authority that insider threat frameworks have long identified as a precursor condition.

Arctic Wolf (2025) observes that traditional business email compromise relied on the exploitation of trust and authority, but that AI now enables threat actors to personalize attacks to the level of an individual's writing quirks, project names, and internal communication style, producing a qualitative leap in credibility. The attack succeeds not by penetrating a network but by satisfying the anthropomorphic heuristics of a human target.

The cost profile reinforces the point. In the 2026 benchmark, the outsmarted insider was the costliest incident type to resolve, at 842,462 dollars per incident, exceeding the negligent insider at 747,107 dollars and the malicious insider at 742,125 dollars, even though credential theft remained the least frequent of the three categories by volume (Ponemon Institute and DTEX Systems, 2026).

5.2 The Cost of Trust Exploitation in the Arup Case

The Arup incident illustrates the operational consequence with unusual clarity. In early 2024 an employee of the engineering firm was deceived into transferring approximately 25 million dollars to fraudulent accounts after participating in a video conference in which all other apparent participants, including senior executives, were AI-generated deepfakes (Wald.ai, 2025). No network intrusion occurred. The entire attack surface was psychological, consisting of the trusted appearance of known colleagues in a familiar communication medium. The incident is categorically an outsmarted insider event, executed through anthropomorphic AI at a fidelity that bypassed ordinary human verification. It is a direct demonstration of the thesis advanced here, that the decisive vulnerability is the human disposition to trust a human-seeming presentation.

5.3 Synthetic Identities and the Ghost Employee Phenomenon

The convergence of generative AI with remote work and digitized hiring has produced a further attack surface that insider threat detection was not designed to monitor. Cheek (2025) describes the emergence of ghost employees, AI-generated personas with fabricated professional histories that are placed within organizational workforces through deepfake credentials and AI-mediated interviews, and cites projections that a substantial share of job candidates globally will be AI-generated fabrications within a few years. The synthetic identity, whether a fully fabricated persona or a real employee impersonated in real time, presents as a trusted insider while operating outside the behavioral and psychological indicators that threat assessment models are calibrated to detect.

This risk is not theoretical. Government and industry documentation has confirmed sustained campaigns by state-sponsored actors, particularly associated with the Democratic People's Republic of Korea, to place AI-assisted fraudulent workers inside technology firms and critical infrastructure operators for the purposes of financial remittance and intelligence access (Wald.ai, 2025; Cheek, 2025). The ghost employee represents the anthropomorphic AI threat vector at its most fully realized, a non-human or misrepresented actor occupying a human organizational role, carrying credentials the organization issued in good faith, and operating through the social trust architecture extended to members.

Sections 4 and 5 together trace the trust pathway from its psychological origin to its operational exploitation, and it is worth consolidating that pathway in the same structured form that the aversion pathway will shortly receive. The elevated trust described above does not remain an attitude. It becomes a set of observable behaviors, and each of those behaviors maps onto the established insider framework. I would argue that this mapping is every bit as direct as the one documented for grievance, with the important difference that the trust pathway operates in the complete absence of hostile motive. Table 4 sets out the principal trust-driven behaviors, the insider typology that each one most closely corresponds to, and the behavioral indicator through which an attentive people leader is most likely to observe it. What the mapping makes plain is that the trust pathway concentrates in the two non-malicious categories, the negligent insider and the outsmarted insider, which is precisely where anthropomorphic design exerts its strongest pull.

Trust-Driven Behavior	Insider Typology	Corresponding Indicator
Disclosing sensitive or regulated material to consumer AI tools believed to be private	Negligent insider	Sensitive data disclosure outside organizational control
Acting on AI output without independent verification	Negligent insider	Automation bias and cognitive offloading
Acting on synthetic voice or video without out-of-band confirmation	Outsmarted insider	Deference to AI-mediated authority
Accepting fluent, context-rich requests that simulate insider knowledge	Outsmarted insider	AI-personalized social engineering and business email compromise
Reluctance to report a suspected anomaly because the interaction felt trustworthy	Negligent insider, shading toward outsmarted	Suppressed reporting and delayed escalation

Table 4. Trust-driven behaviors mapped to insider typologies and behavioral indicators. Synthesized from the trust and exploitation mechanisms developed in Sections 4 and 5.

6. The Aversion Pathway, AI-Induced Grievance and Insider Risk

6.1 Two Poles of the Same Relational Shift

The argument to this point has followed a single pathway, the one that runs through trust. Anthropomorphic design invites over-trust, and over-trust produces the negligent disclosure and the outsmarted compliance that define two of the three insider typologies. There is, however, a second and opposite pathway that the same relational shift produces, and it terminates in the third typology, the disgruntled and malicious insider. Where the first pathway is driven by what employees come to feel toward human-seeming systems, the second is driven by what they come to resent. Alqasir (2025) frames both as native human responses to AI, pairing the attachment and parasocial bonding that the trust literature documents with the algorithm aversion that a parallel literature records. Treating the two together yields a sharper claim than either supports alone. Artificial intelligence is reshaping insider risk from both directions at once, through the affection it cultivates and through the aversion it provokes.

6.2 The Documented Rise of AI Anger

Aversion is no longer a fringe sentiment, and it is intensifying. In successive annual surveys of Americans aged 14 to 29, the Walton Family Foundation, GSV Ventures, and Gallup found that the share of Gen Z who report feeling angry about AI rose by nine percentage points in a single year to 31 percent, while excitement fell by fourteen points to 22 percent and hopefulness declined to 18 percent (Walton Family Foundation et al., 2026). Anger and anxiety, not excitement and hope, are now the dominant emotions in this cohort, even as usage holds roughly steady, which indicates that the negative shift is not a function of unfamiliarity. Broader polling is consistent, with reporting that registered voters hold negative views of AI by a wide margin over positive ones (Angelo, 2026). The significance for insider risk is that the workforce entering organizations now carries a markedly more hostile disposition toward the technology than the workforce of even two years ago.

6.3 From Grievance to Action, the Psychological Mechanism

The path from a negative attitude to a harmful act is well charted in organizational psychology, and recent work has extended it specifically to AI. The mechanism runs through job insecurity. Shoss (2017) established technological change as an antecedent of perceived job insecurity and job insecurity in turn as a predictor of withdrawal, resistance, and counterproductive work behavior. Applying this framework to the present moment, a three-wave study of office employees found that AI awareness produces deviant behavior, with job insecurity serving as the mediating mechanism, such that employees who perceive their roles as replaceable develop negative perceptions that translate into deviance and reduced performance (Chung et al., 2025). Conservation of Resources theory supplies the motivational logic. When people perceive a threat to a valued resource such as job stability, they take defensive or retaliatory action to protect what remains (Hobfoll, 1989). The technostress literature converges on the same point,

linking AI-induced job insecurity to heightened stress, active resistance to the technology, and reduced organizational commitment (Brougham and Haar, 2018; Califf et al., 2020). Anger at AI, on this account, is not an inert feeling. It is a stressor that predictably converts into behavior, and some of that behavior is indistinguishable from insider activity.

6.4 Sabotage as Insider Behavior, the Empirical Signal

That conversion is now visible in the field. A 2026 survey of 2,400 knowledge workers across the United States, the United Kingdom, and Europe, including 1,200 senior executives, found that 29 percent of employees admitted to actively undermining their organization's AI strategy, a figure that rose to 44 percent among Gen Z (Writer and Workplace Intelligence, 2026). The reported forms of that sabotage are, almost item for item, the behaviors catalogued in the insider typologies of Section 3. Workers described entering proprietary information into unapproved public AI tools, using unsanctioned applications, deliberately generating low-quality output to discredit a mandated system, refusing to use required tools, and tampering with performance metrics. Table 5 maps these aversion-driven behaviors onto the insider framework.

Aversion-Driven Behavior	Insider Typology	Corresponding Indicator
Entering proprietary data into public or unapproved AI tools	Negligent insider, shading toward malicious	Sensitive data exfiltration outside organizational control
Refusing mandated tools; organizing human-only workflows	Disgruntled insider, resistance posture	Disengagement and policy non-compliance
Deliberately degrading output to discredit a tool	Malicious insider, low-level sabotage	Intentional performance manipulation
Tampering with performance metrics or reviews	Malicious insider	Falsification of organizational records
Threats or hostility following AI layoffs	Grievance escalation	Concerning communications and threat behavior

Table 5. Aversion-driven behaviors mapped to insider typologies and behavioral indicators. Behavior categories drawn from Writer and Workplace Intelligence (2026).

The surrounding picture of unsanctioned use confirms the scale of the exposure. Industry analysis of the same period reports that a substantial majority of employees bypass official channels to use unsanctioned AI tools, that the typical enterprise records hundreds of AI-related data policy violations each month, and that the majority of organizations worried about sensitive data leaking through generative AI still lack a specific strategy to prevent it (DTEX Systems, 2026). The disgruntled insider, in other words, no longer needs to exfiltrate data through a covert channel. An ordinary consumer chatbot, used in anger or in protest, accomplishes the same result. Leadership itself now treats this resistance as material, with 76 percent of executives describing employee sabotage as a serious threat to their company's future and 67 percent believing their organization has already suffered a data leak or breach through an employee's use of an unapproved AI tool (Writer and Workplace Intelligence, 2026).

6.5 The True Locus of the Grievance

Two cautions sharpen the analysis rather than weaken it. First, much of the most striking evidence is self-reported and, in the case of the workforce survey, produced by an enterprise AI vendor, so the precise magnitudes warrant treatment as indicative rather than definitive. Second, and more important for practice, not all of this resistance is irrational hostility toward the technology. A meaningful share of the employees who admitted to sabotage cited poor organizational AI strategy rather than personal job fear as their reason, and a majority of executives in the same survey conceded that their AI strategy was more presentation than substance (Writer and Workplace Intelligence, 2026). The grievance that converts most reliably into insider behavior is therefore less an objection to AI as such and more a reaction to how leadership imposes it, through mandates issued without voice, through layoffs, and through messaging that signals to employees that they are regarded as expendable. This is the territory of organizational justice, and it carries a constructive implication. Because the decisive variable is a perception of unfair treatment rather than the technology itself, it is a variable that leaders can directly influence, which positions the people leader, the subject of the next section, as the primary control on the aversion pathway.

7. Recognizing the Behaviors, A Guide for People Leaders

Because the vulnerability is behavioral, the earliest and most reliable indicators are behavioral as well, and they are observable by attentive people leaders who know what to look for. The literature reviewed above, read alongside established practice in behavioral threat assessment, suggests a set of recognizable indicators that an individual or team has crossed from appropriate use into anthropomorphic over-trust. These indicators are not in themselves evidence of wrongdoing. They are signals that the conditions for negligent disclosure or successful deception are present and rising.

Relational language toward tools. Employees who refer to an AI system by a personal name or gendered pronoun, describe it as a colleague, or attribute intentions and feelings to it have adopted a mental model that lowers scrutiny.

Sensitive disclosure to unsanctioned tools. Pasting confidential, regulated, or proprietary material into consumer AI tools, often justified by the belief that the tool is private or secure, is a direct enactment of the disclosure risk.

Acceptance without verification. A pattern of acting on AI output without independent checking, and difficulty explaining the reasoning behind an AI-assisted decision, indicates automation bias and cognitive offloading.

Deference to AI-mediated authority. Acting on instructions delivered by voice or video without out-of-band confirmation, especially for financial or access-related requests, is the precise behavior the Arup-type attack exploits.

> **Attachment and distress at change.** Visible reluctance to change tools, or distress when a familiar system is updated or withdrawn, signals a parasocial bond that has displaced critical distance.
>
> **Suppressed reporting.** A reduced willingness to flag an anomaly because the interaction felt trustworthy reflects the seductive quality of human-like presentation overriding professional skepticism.

Recognition is necessary but not sufficient. People leaders must also be able to convert recognition into reporting, and this is where many programs fail. A useful diagnostic frame distinguishes three postures that employees adopt toward reporting a concern, the reflexive refusal that begins with no because, the hesitant acknowledgment that begins with yes but, and the conditional willingness that begins with yes if. The leader's task is to identify and remove the specific barrier behind each posture, the uncertainty about whether a soft signal warrants escalation, the fear of being wrong about a tool that feels benign, or the absence of a clear and low-friction channel. Lowering these barriers is especially important for anthropomorphic AI incidents, where the deception is engineered to feel like an ordinary, trustworthy interaction and therefore to suppress the instinct to report.

8. Organizational Mitigation

8.1 Governing Anthropomorphic AI as a Human-Factors Risk

The first move is conceptual. Organizations should treat anthropomorphic design not as a neutral feature of the tools they deploy but as an active human-factors risk variable that belongs in their threat modeling. This entails inventorying anthropomorphic AI deployments, the assistants, agents, and avatars embedded across communication and decision support, and documenting where each is positioned to elicit trust and disclosure. The governance posture that security practitioners increasingly recommend for AI tools, including least-privilege access, rigorous authentication, and continuous monitoring of how sensitive data flows into and out of these systems, applies with equal force to the human use of the tools as to the tools themselves (ISACA, 2024; Citrix, 2025).

8.2 Awareness Training and Verification Protocols

Security awareness training built for an earlier threat environment, oriented toward recognizing generic phishing, is insufficient against AI-personalized deception. Training should be revised to address the specific signatures of AI-enabled social engineering, including synthetic voice and video, AI-personalized business email compromise, and synthetic colleague impersonation, and to establish the habit that high-stakes requests are verified through an independent channel regardless of how convincing the originating communication appears. The single most valuable behavioral control against the Arup-type attack is a simple, non-negotiable out-of-band verification protocol for financial transfers and access changes, one that does not depend on recognizing that a deepfake is a deepfake, because by design it cannot be recognized.

8.3 Design and Language Intervention

Where an organization controls the design of its internal AI tooling, it can reduce anthropomorphic over-trust at the source. Practical measures include avoiding gendered names and voices and first-person framing that implies an inner life, clearly and persistently labeling AI-generated output as such, and incorporating cognitive forcing functions that require a user to engage their own reasoning before accepting a consequential recommendation (Buçinca et al., 2021). The same principle extends to internal communication and policy. Adopting accurate, de-anthropomorphized language, processing rather than thinking, unsupported output rather than hallucination, reduces the inflated intuitions about capability and trustworthiness that the field's standard vocabulary otherwise propagates (Salles et al., 2020).

8.4 Calibrated Trust and Reporting Culture

The objective is not to eliminate trust in AI tools, which would be neither realistic nor productive, but to calibrate it, aligning reliance with the system's actual reliability and limits (Alqasir, 2025). Calibrated-trust training teaches employees where these systems are competent and where they fail, so that reliance tracks capability rather than the warmth of the interface. This technical calibration must be paired with a reporting culture that lowers the barriers identified in Section 7, so that an employee who suspects a deepfake, a manipulated request, or an over-disclosure incident encounters a clear, low-friction, and psychologically safe channel for raising it. In a threat environment engineered to feel trustworthy, the willingness to report a quiet unease is among the most valuable behavioral controls an organization can cultivate.

9. Limitations and Future Research

Several limitations constrain the conclusions offered here and define an agenda for further work. First, much of the empirical literature on anthropomorphism and trust relies on controlled experiments with non-representative samples. Longitudinal studies of how sustained exposure to anthropomorphic AI affects employee security behavior, trust calibration, and policy compliance in live organizational settings do not yet exist, and are needed. Second, the academic literature on AI trust and the practitioner literature on behavioral threat assessment remain largely parallel. Research that bridges established behavioral indicators for human insiders with the behavioral signatures of AI-enabled deception would materially strengthen detection. Third, the outsmarted insider, the category most directly exposed to anthropomorphic social engineering and the costliest to resolve on a per-incident basis, still lacks the specific detection signatures, training protocols, and verification controls that its evolved threat profile now demands. Fourth, industry benchmarks are annual and necessarily trail real-time capability, so the gap between current attacker sophistication and typical organizational defense is likely wider than published figures reflect. Finally, the present paper deliberately brackets a significant adjacent question, whether autonomous AI agents themselves exhibit behaviors that match the insider threat profile. That question warrants its own sustained treatment and is reserved for a separate paper.

10. Conclusion

Anthropomorphic AI design is not a passive environmental feature. It is an active risk variable, and the risk it creates is fundamentally human. The same design choices that make these systems more effective and more pleasant to use, the names and voices, the simulated memory and empathy, the fluent and contextually apt responses, also make them more effective vectors for the exploitation of trust. The vulnerability they exploit is not a defect peculiar to careless employees. It is a stable feature of human social cognition, cultivated from childhood through anthropomorphic media and reinforced by a technical vocabulary that imports mind where there is only computation.

For insider risk, three implications follow. The disposition to humanize machines lowers the verification and skepticism on which security behavior depends, and it does so most powerfully among the least expert users. Adversaries are already weaponizing that disposition through deepfakes, AI-personalized impersonation, and synthetic identities that defeat human verification by design. And the most tractable defenses are correspondingly human, the trained recognition of behavioral indicators by attentive people leaders, out-of-band verification that does not depend on detecting a forgery, the calibration of trust to actual reliability, accurate and de-anthropomorphized language, and a reporting culture that lowers the barriers to surfacing a quiet unease. Organizations that continue to treat insider risk as a purely technical or purely human-malice problem will remain exposed to a threat that operates precisely in the space between the human mind and the human-seeming machine.

References

Airenti, G. (2018). The development of anthropomorphism in interaction: Intersubjectivity, imagination, and theory of mind. Frontiers in Psychology, 9, 468. https://doi.org/10.3389/fpsyg.2018.00468

Alqasir, A. (2025). The relational shift: Why we need "AI Psychology" now as a core field. Journal of Psychology and AI, 1(1), 2573928. https://doi.org/10.1080/29974100.2025.2573928

Angelo, J. (2026, April 8). Gen Z workers are so fearful AI will take their job they're intentionally sabotaging their company's AI rollout. Fortune. https://fortune.com/2026/04/08/gen-z-workers-sabotage-ai-rollout-backlash/

Arctic Wolf. (2025, November 25). How to combat AI-enhanced social engineering attacks. https://arcticwolf.com/resources/blog/how-to-combat-ai-enhanced-social-engineering-attacks/

Brougham, D., & Haar, J. (2018). Smart technology, artificial intelligence, robotics, and algorithms (STARA): Employees' perceptions of our future workplace. Journal of Management & Organization, 24(2), 239--257. https://doi.org/10.1017/jmo.2016.55

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce over-reliance on AI in AI-assisted decision making. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1), 1--21. https://doi.org/10.1145/3449287

Califf, C. B., Sarker, S., & Sarker, S. (2020). The bright and dark sides of technostress: A mixed-methods study involving healthcare IT. MIS Quarterly, 44(2), 809--856. https://doi.org/10.25300/MISQ/2020/14818

Cheek, P. (2025, September 10). An emerging cyber security threat: The growing threat of AI employees, insider social engineering, and corporate exploitation. https://www.paulcheek.com/articles/an-emerging-cyber-security-threat

Cheng, M., DeVrio, A., Egede, L., Blodgett, S. L., & Olteanu, A. (2024). "I am the one and only, your cyber BFF": Understanding the impact of GenAI requires understanding the impact of anthropomorphic AI [Preprint]. arXiv. https://arxiv.org/abs/2410.08526

Chung, Y. W., Im, S., Kim, J. E., & Yun, J. K. (2025). Artificial intelligence awareness, career resilience, job insecurity and behavioural outcomes. Australian Journal of Psychology, 77(1), Article 2559910. https://doi.org/10.1080/00049530.2025.2559910

Citrix. (2025, August 4). AI agents are the new insider threat: Secure them like human workers. Citrix Blogs. https://www.citrix.com/blogs/2025/08/04/ai-agents-are-the-new-insider-threat-secure-them-like-human-workers/

Cybersecurity and Infrastructure Security Agency. (2025). Insider threat mitigation guide. U.S. Department of Homeland Security. https://www.cisa.gov/resources-tools/resources/insider-threat-mitigation-guide

Cybersecurity and Infrastructure Security Agency. (2026). Multi-disciplinary insider threat management team guidance. U.S. Department of Homeland Security. https://www.cisa.gov

Cybersecurity and Infrastructure Security Agency, National Security Agency, & Federal Bureau of Investigation. (2025, May 22). AI data security guidance. U.S. Department of Homeland Security. https://www.cisa.gov

Das, B. C., Sartaz, M. S., Reza, S. A., Hossain, A., Nasiruddin, M., Bishnu, K. K., Sharmeen Shatyi, S., & Khan, M. A. (2025). AI-driven cybersecurity threat detection: Building resilient defense systems using predictive analytics. International Journal of Basic and Applied Sciences, 14(4), 33--45. https://doi.org/10.14419/hysdg957

Deshpande, A., Rajpurohit, T., Narasimhan, K., & Kalyan, A. (2023). Anthropomorphization of AI: Opportunities and risks. In Proceedings of the Natural Legal Language Processing Workshop 2023 (pp. 1--7). Association for Computational Linguistics. https://aclanthology.org/2023.nllp-1.1/

Dietz, G., Outa, J., Lowe, L., Landay, J. A., & Gweon, H. (2023). Theory of AI mind: How adults and children reason about the "mental states" of conversational AI. In Proceedings of the 45th Annual Meeting of the Cognitive Science Society (CogSci 2023) (Vol. 45). Cognitive Science Society.

DTEX Systems. (2026, January 13). Moving from insider risk to insider threat: Disgruntled employees and looming layoffs. https://www.dtex.ai/blog/moving-from-insider-risk-to-insider-threat-disgruntled-employees-and-looming-layoffs/

Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. Societies, 15(1), 6. https://doi.org/10.3390/soc15010006

Gong, C., & Yang, Y. (2024). Google effects on memory: A meta-analytical review of the media effects of intensive Internet search behavior. Frontiers in Public Health, 12, 1332030. https://doi.org/10.3389/fpubh.2024.1332030

Guingrich, R. E., & Graziano, M. S. A. (2024). Ascribing consciousness to artificial intelligence: Human--AI interaction and its carry-over effects on human--human interaction. Frontiers in Psychology, 15, 1322781. https://doi.org/10.3389/fpsyg.2024.1322781

Hobfoll, S. E. (1989). Conservation of resources: A new attempt at conceptualizing stress. American Psychologist, 44(3), 513--524. https://doi.org/10.1037/0003-066X.44.3.513

IBM Security. (2024). The dangers of anthropomorphizing AI: An infosec perspective. IBM Think Insights. https://www.ibm.com/think/insights/anthropomorphizing-ai-danger-infosec-perspective

ISACA. (2024). AI security risk and best practices: Industry news. https://www.isaca.org/resources/news-and-trends/industry-news/2024/ai-security-risk-and-best-practices

Kotb, H. M., Gaber, T., AlJanah, S., Zawbaa, H. M., & Alkhathami, M. (2025). A novel deep synthesis-based insider intrusion detection (DS-IID) model for malicious insiders and AI-generated threats. Scientific Reports, 15. https://doi.org/10.1038/s41598-024-84673-w

PMC / National Library of Medicine. (2025). The benefits and dangers of anthropomorphic conversational agents. PubMed Central. https://pmc.ncbi.nlm.nih.gov/articles/PMC12146756/

Ponemon Institute / DTEX Systems. (2023). Cost of insider risks: Global report 2023. https://ponemonsullivanreport.com/2023/10/cost-of-insider-risks-global-report-2023/

Ponemon Institute / DTEX Systems. (2025). 2025 cost of insider risks global report. https://ponemon.dtexsystems.com/

Ponemon Institute / DTEX Systems. (2026). 2026 cost of insider risks global report. https://ponemon.dtex.ai/

Reani, M., Crockett, K., & Sherrat, S. (2025). Anthropomorphism on risk perception: The role of trust and domain knowledge in decision-support AI [Preprint]. arXiv. https://arxiv.org/abs/2602.13625

Salles, A., Evers, K., & Farisco, M. (2020). Anthropomorphism in AI. AJOB Neuroscience, 11(2), 88--95. https://doi.org/10.1080/21507740.2020.1740350

ScienceDirect / Computers in Human Behavior. (2025). Exploring dimensions of perceived anthropomorphism in conversational AI: Implications for human identity threat and dehumanization. Computers in Human Behavior Open, 2(3). https://doi.org/10.1016/j.chbopen.2025.100763

Shoss, M. K. (2017). Job insecurity: An integrative review and agenda for future research. Journal of Management, 43(6), 1911--1939. https://doi.org/10.1177/0149206317691574

Springer Nature / AI and Ethics. (2024). Anthropomorphism in AI: Hype and fallacy. AI and Ethics. https://doi.org/10.1007/s43681-024-00419-4

Wald.ai. (2025, September 30). Gen AI security breaches timeline (2023--2025): Recurring mistakes are the real threat. https://wald.ai/blog/gen-ai-security-breaches-timeline-20232025-recurring-mistakes-are-the-real-threat

Walton Family Foundation, GSV Ventures, & Gallup. (2026). Voices of Gen Z: The AI paradox. https://www.waltonfamilyfoundation.org/

Writer, & Workplace Intelligence. (2026, April 7). Enterprise AI adoption in 2026: Why 79% face challenges despite high investment. Writer. https://writer.com/blog/enterprise-ai-adoption-2026/

Xiao, Y., & Ng, L. H. X. (2025). Humanizing machines: Rethinking LLM anthropomorphism through a multi-level framework of design [Preprint]. arXiv. https://arxiv.org/abs/2508.17573

Zvelo. (2024, April 24). The role of AI in social engineering. https://zvelo.com/the-role-of-ai-in-social-engineering/

About the Author

Dr. Joshua L. Barrett, DSL, is a human-centric insider risk researcher and practitioner who studies all things insider risk and threat through a human lens, with specific focus on leaders, security culture, and reporting cultures. A retired U.S. Marine Corps Chief Warrant Officer with a counterintelligence and human intelligence background spanning the Intelligence Community, the Department of Defense, and the financial services sector, he holds a Doctor of Strategic Leadership and is the founder of Dolos Imperative LLC.

About Dolos Imperative

Dolos Imperative LLC delivers human elements security consulting, research, and training, helping organizations understand and manage the human dimension of insider risk. Learn more at dolosimperative.com.