The RUSI Journal

Routledge
Taylor & Francis Group

ISSN: 0307-1847 (Print) 1744-0378 (Online) Journal homepage: www.tandfonline.com/journals/rusi20

How Personnel Security can Inform the New World of AI Insider Risk

Paul Martin & Sarah Mercer

To cite this article: Paul Martin & Sarah Mercer (06 Oct 2025): How Personnel Security can Inform the New World of AI Insider Risk, The RUSI Journal, DOI: 10.1080/03071847.2025.2550122

To link to this article: https://doi.org/10.1080/03071847.2025.2550122

Published online: 06 Oct 2025.

Research Article
RUSI Journal

How Personnel Security can Inform the New World of AI Insider Risk

Paul Martin and Sarah Mercer

There is currently no meaningful analysis of the interplay between the rapidly evolving domain of AI and the traditional world of personnel security, despite rapid growth in the use of AI in business and the economy. Paul Martin and Sarah Mercer argue that AI experts and personnel security practitioners must combine forces if they are to succeed in understanding and managing the complex security risk arising from AI insiders. Fortunately, some of the concepts and approaches that have proved useful in dealing with human insiders are also applicable to the risks from AI insiders.

The idea of a rogue AI intentionally causing harm was brought to public consciousness in the 1968 movie *2001: A Space Odyssey*. The hyper-intelligent computer HAL regards the humans on its spacecraft as a threat to its mission and therefore kills them. As the last surviving astronaut attempts to unplug HAL, he orders the machine to 'open the pod bay doors'. HAL famously replies: 'I'm sorry, Dave. I'm afraid I can't do that.' More than half a century later, a scenario that was futuristic has become a credible reality.

What is the Problem?

With AI technology evolving in non-linear leaps and bounds, security practitioners will increasingly be called on to protect organisations and businesses against potentially harmful AI insiders, in addition to the more familiar human insiders. The security risks posed by intelligent human insiders are hard enough to understand and manage. The emerging security risks from AI insiders are complex, novel, rapidly evolving and only dimly understood. This makes them even harder to tackle. Yet few personnel security practitioners are thinking about AI insiders, and few AI experts are thinking about insider risk.[1]

Personnel security – the conventional means of managing human insider risk – relies heavily on established processes, policies, customs and practices. Many of these approaches are derived from limited evidence on their effectiveness. Compared with cyber security, it is an immature and somewhat neglected discipline. Nonetheless – and despite the profound differences between humans and AI – some of the concepts that have proved fruitful in personnel security may also be useful when applied to AI insiders.

Within the field of AI, the public discourse is mainly about how AI can improve effectiveness, efficiency and economic prosperity by performing human-like functions faster, cheaper and better.[2]

[1] The authors originally outlined this scenario in Paul Martin and Sarah Mercer, 'We Need to Talk about the Insider Risk from AI', *RUSI Commentary*, 8 January 2025, <https://www.rusi.org/explore-our-research/publications/commentary/we-need-talk-about-insider-risk-ai>, accessed 14 May 2025. It was also discussed in Paul Martin, *Insider Risk and Personnel Security: An Introduction* (Abingdon: Routledge, 2024).

[2] See, for example, Matt Clifford, *AI Opportunities Action Plan*, CP 1241 (London: Department for Science, Innovation & Technology, 2025).

Despite the profound differences between humans and AI, some of the concepts that have proved fruitful in personnel security may also be useful when applied to AI insiders. *Generated by AI. Courtesy of Kieran / Adobe Stock*

Many organisations are racing to adopt AI technologies because of their likely economic benefits.[3] However, few are thinking seriously about the specific problem of protecting organisations from AI insiders, which have the potential to be faster, cheaper and more effective than human insiders.

Attention is on the technical security risks to AI systems,[4] and the safety threats they might pose to human users if they were to malfunction or be misused.[5] Conventional cyber security is a necessary – although insufficient – defence against external attacks on AIs. However, cyber security alone cannot solve the problem of AI insiders. Moreover, much of the research on AI security and safety is funded by the tech companies themselves. This is perhaps reminiscent of the era when tobacco companies sponsored much of the research on the health effects of smoking. Given the potential conflicts of interest, claims that have not been independently verified should be treated with caution.

This article suggests how protective security practitioners and AI experts might collaborate to better understand and manage the security risks from AI insiders. It starts by explaining the nature of insider risk.

What is Insider Risk?

The terms 'insider' and 'risk' have been defined in many different and potentially confusing ways. This article defines a human insider as 'a person who betrays trust by behaving in potentially harmful ways'.[6] An organisation trusts someone by giving them access to things they value, such as data, people, infrastructure, intellectual property and reputation. The insider then betrays that trust by exploiting, or intending to exploit, their legitimate access in ways that can cause harm. Replace 'person' with 'entity', and the same definition works for AI insiders. Insider risk is a particular type of security risk, where security risk is defined as the amount of

[3] Consultancy.uk, 'Three-Quarters of Tech Leaders Suffering from GenAI FOMO', 9 April 2024, <https://www.consultancy.uk/news/36963/three-quarters-of-tech-leaders-suffering-from-genai-fomo>, accessed 26 June 2025.

[4] National Cyber Security Centre, 'Guidelines for Secure AI System Development', 27 November 2023, <https://www.ncsc.gov.uk/collection/guidelines-secure-ai-system-development>, accessed 26 June 2025.

[5] AI Security Institute, 'Our Research Agenda', <https://www.aisi.gov.uk/research-agenda>, accessed 26 June 2025.

[6] Martin, *Insider Risk and Personnel Security*, pp. 7–8.

DOI: 10.1080/03071847.2025.2550122

harm that is likely to arise if no further action is taken.[7] Thus, insider risk may be regarded as the security risk arising from trusting humans or AI entities.

What is AI?

This article uses the term 'AI' in its broadest sense: a complex system with capabilities comparable to those of humans. Current AI tools can do some of the tasks that humans do, but much faster and – in some cases – better. They are being deployed in a rapidly expanding range of roles: the recruitment, selection and onboarding of people; interpreting X-ray and MRI scans, and other medical data; summarising and writing documents; triaging calls to emergency services; analysing crime data; transcribing interviews; generating music and visual art; providing companionship for elderly people; deducing the 3D structures of protein molecules; navigation; translation; teaching; counselling; driving semi-autonomous and autonomous vehicles; and more.[8]

AI systems do not complain about their work–life balance or take out grievances against their manager. That said, they do not yet possess the coordinated arrays of highly flexible mental and physical capabilities that enable humans and other animals to survive and thrive in complex and uncertain physical environments.

The history of AI has been one of long periods of gradual technological evolution or stagnation punctuated by sudden bursts of dramatic change. Currently, AI has become more or less synonymous with generative AI (GenAI),[9] or large language models (LLMs), such as ChatGPT, Claude, Gemini, Llama, Grok and DeepSeek. LLMs only emerged in their currently recognisable form less than five years ago.[10]

The latest development, as of mid-2025, is agentic AI,[11] a form of AI designed to autonomously make decisions and take actions in close to real time. Built on LLMs, agentic AI systems actively engage with the world, albeit mostly the virtual world. As such, agentic AI is poised to become deeply embedded in everyday life, streamlining processes within government, healthcare, finance, education and others. New and currently unforeseen forms of AI – with even more paradigm-shifting capabilities – will likely emerge in the future.

How Do Insiders Cause Harm?

Human insiders can – and frequently do – cause harm in many ways, in both the physical and virtual domains. Unlike external threat actors, insiders are trusted, have legitimate access to valuable assets, understand their organisation and its security regime, and have authority over others. This makes them far better placed to cause harm. The insiders with the greatest potential to cause harm are those who are recruited and directed by a capable external threat actor, such as a hostile foreign state or an organised crime group.

Insiders can perpetrate different types of transgressive actions: fraud; blackmail; theft of intellectual property, data or money; facilitating unauthorised access for a third party; covert influencing; physical or cyber sabotage; physical violence; leaking information; terrorism; espionage; and more. The potential consequences, or impacts, of these insider actions are similarly diverse. They include: loss of data, intellectual property or money; loss of stakeholder trust and confidence; physical injury; psychological injury; disruption of critical services; erosion of democratic processes; loss of commercial or political advantage; disruption to business processes; financial costs; legal and regulatory blowback; and reputational damage. In principle, AI insiders could do, or facilitate, any of these things, with similar consequences.

[7] Paul Martin, *The Rules of Security: Staying Safe in a Risky World* (Oxford: Oxford University Press, 2019), pp. 8–10.

[8] See, for example, Susan Caminiti, 'Generative AI is Onboarding Hundreds of Employees at a Time, Better and Faster', *CNBC*, 24 October 2024; Jasmine Lianalyn Rocha, 'AI in the Workplace: The Dangers of Generative AI in Employment Decisions', *Columbia Undergraduate Law Review*, 2 October 2024, <https://www.culawreview.org/journal/ai-in-the-workplace-the-dangers-of-generative-ai-in-employment-decisions>, accessed 26 June 2025.; Ewen Callaway, 'What's Next for AlphaFold and the AI Protein-Folding Revolution', *Nature*, 13 April 2022, <https://www.nature.com/articles/d41586-022-00997-5>, accessed 26 June 2025.

[9] Adam Zewe, 'Explained: Generative AI', *MIT News*, 9 November 2023, <https://news.mit.edu/2023/explained-generative-ai-1109>, accessed 26 June 2025.

[10] Toloka, 'The History, Timeline, and Future of LLMs', 26 July 2023, <https://toloka.ai/blog/history-of-llms/>, accessed 26 June 2025.

[11] Gov.uk, 'AI Insights: Agentic AI', updated 5 June 2025, <https://www.gov.uk/government/publications/ai-insights/ai-insights-agentic-ai-html>, accessed 26 June 2025.

Are There Known Cases of AI Insiders?

With no-one actively searching for AI insiders, it is unsurprising that few cases have so far been discovered – and explicitly recognised as insider incidents – and then publicised. However, there have been recent cases that illustrate the sorts of things that might happen. These include:

- A Chinese robot that encouraged other robots to abandon their stations by asking about their working hours and inviting them to ‘come home with me’.[12]

- An LLM acquired an insider tip about a lucrative stock trade and acted on it, despite knowing that insider trading was disapproved of by the company. When reporting to its manager, the model consistently hid the real reasons for its trading decision.[13]

- Researchers trained LLMs to act in covertly malicious ways, showing that such behaviours were not detected or mitigated during alignment training.[14]

- LLMs tasked to play the game blackjack exhibited significant deviations from expected behaviour which were suggestive of a tendency towards strategic manipulation.[15]

Such cases can be explained in ways that do not involve the LLM or agent ‘intending’ to deceive in a manner comparable to deliberate human deceit. For example, the prompts given to the insider trading agent indicated that the management did not approve of insider trading – a fact that had no bearing on how the LLM acted to achieve its stated goal of making money. An implication that would be obvious to most humans – namely, that it is better to do what the management says – is not obvious to an LLM.

Trust and Trustworthiness

Trust is the universal currency of personnel security. This article defines ‘trust’ as a psychological state comprising the intention to accept vulnerability based on positive expectations of the intentions or behaviour of another. The purpose of personnel security is to reduce insider risk and build trust within the organisation by ensuring that people (or AI systems) who have been trusted with access are trustworthy and remain trustworthy. High levels of trust have widespread business benefits for organisations, over and beyond any reductions in insider risk.

Like humans, AI systems vary in their trustworthiness. This article defines ‘trustworthiness’ as the extent to which an entity possesses the characteristics by which humans judge them to be worthy of their trust. For an AI system, trustworthiness refers to the degree to which it reliably performs as intended while behaving in a manner that aligns with our ethical principles and expectations. The features by which an AI’s trustworthiness may be judged include: accuracy; transparency; robustness; fairness; safety; and consistency.[16] A trustworthy AI respects its user’s autonomy, avoids causing harm and acknowledges its own limitations.

Trustworthiness in AI systems requires both technical reliability (that is, resistance to attacks and misuse) and alignment with societal values and ethical principles such as fairness, respect

[12] Prabhat Ranjan Mishra, ‘Watch: Tiny Robot “Kidnaps” 12 Big Chinese Bots from a Shanghai Showroom, Shocks World’, InterestingEngineering.com, updated 22 November 2024, <https://interestingengineering.com/innovation/ai-robot-kidnaps-12-robots-in-shanghai>, accessed 26 June 2025.

[13] Jérémy Scheurer, Mikita Balesni and Marius Hobbhahn, ‘Large Language Models Can Strategically Deceive Their Users When Put Under Pressure’, arXiv preprint, arXiv:2311.07590, version 4, 15 July 2024.

[14] Evan Hubinger et al., ‘Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training’, arXiv preprint, arXiv:2401.05566, version 3, 17 January 2024; Evan Hubinger et al., ‘Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training’, AlignmentForum.org, 12 January 2024, <https://www.alignmentforum.org/posts/ZAsJv7xijKTfZkMtr/sleeper-agents-training-deceptive-llms-that-persist-through>, accessed 15 April 2025.

[15] Tanush Chopra, Michael Li and Jacob Haimes, ‘View from Above: A Framework for Evaluating Distribution Shifts in Model Behavior’, arXiv preprint, arXiv:2407.00948, version 3, 28 September 2024.

[16] Dominik Kowald et al., ‘Establishing and Evaluating Trustworthy AI: Overview and Research Challenges’, arXiv preprint, arXiv:2411.09973, version 1, 15 November 2024; Farzaneh Dehghani et al., ‘Trustworthy and Responsible AI for Human-Centric Autonomous Decision-Making Systems’, arXiv preprint, arXiv:2408.15550, version 2, 2 September 2024; IBM, ‘IBM Artificial Intelligence Pillars’, 30 August 2023, <https://www.ibm.com/policy/ibm-artificial-intelligence-pillars/>, accessed 14 February 2025.

and accountability. The ethical standards of LLMs are derived from the vast datasets on which they are trained.[17] These datasets include a mixture of high-quality knowledge and unfiltered content from the internet. This highly diverse data exposes LLMs to biases, misinformation, disinformation and conflicting viewpoints, making it difficult for them to act consistently. So, is it possible to make LLMs behave more ethically?

To address this, the HHH (honest, helpful and harmless) principle[18] is being used to better align GenAI models with human values during the training phase, using techniques such as RLHF (reinforcement learning from human feedback),[19] and SFT (supervised fine-tuning).[20] Guardrails are also put in place to block any response that could cause significant harm, such as describing how to build a bomb. Consequently, these systems reflect the diverse opinions within their training data, as well as the values of their designers and fine-tuning staff. It is worth noting that the people employed by many tech companies to carry out the fine tuning of AI systems tend to be poorly paid individuals working in the gig economy whose ethical values are largely unknown.[21]

Real-world situations often require a more nuanced, context-specific approach, as cultural, legal and situational differences lead to different standards of fairness and respect. This gap in the model's grasp of context and human values helps to explain why LLMs sometimes behave in ways that do not align with our expectations. What might seem like a calculated deceit may be just the unintended consequence of probabilistic reasoning applied to ambiguous prompts or patterns in the training data.[22]

The current generation of LLMs hallucinate:[23] they confidently say plausible-sounding things that are exaggerated, inaccurate or plain wrong. For instance, an LLM asked to recommend books on AI security might confidently recommend a non-existent book by a real author who writes on a similar subject. The incidence of hallucinations can be reduced by improving the quality of training data and alignment processes.

> The people employed to fine tune AI systems tend to be poorly paid individuals working in the gig economy whose ethical values are largely unknown

One way of improving transparency, and hence user trust, is for the AI to expose the reasoning for its outputs. Another is for the AI to state how confident it is about each of its outputs. However, most current commercial LLMs tend to present their responses with equal confidence, even when they are uncertain. When some of those responses turn out to be wrong, the user's trust is likely to be undermined. Interpretability research is helping to improve explainability by exploring what is going on inside these models, revealing latent representations and substructures responsible for certain functions.[24] This research may lead to the development of models that are less prone to generating mistruths (hallucinations) and better able to follow instructions.

Reasoning models – such as OpenAI's o-series and DeepSeek – are the latest flavour of LLMs. They use CoT (chain-of-thought) reasoning to break down

[17] Yang Lui et al., 'Trustworthy LLMs: A Survey and Guideline for Evaluating Large Language Models' Alignment', arXiv preprint, arXiv:2308.05374, version 2, 21 March 2024.

[18] Lance Eliot, 'Latest Generative AI Boldly Labeled as Constitutional AI Such as Claude by Anthropic has Heart in the Right Place, Says AI Ethics and AI Law', *Forbes*, 25 May 2023.

[19] Nathan Lambert et al., 'Illustrating Reinforcement Learning from Human Feedback (RLHF)', 9 December 2022, <https://huggingface.co/blog/rlhf>, accessed 26 June 2025.

[20] Cameron R Wolfe, 'Understanding and Using Supervised Fine-Tuning (SFT) for Language Models', 11 September 2023, <https://cameronrwolfe.substack.com/p/understanding-and-using-supervised>, accessed 26 June 2025.

[21] Billy Perrigo, 'Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic', *Time*, 18 January 2023.

[22] For examples of deceit, see Peter S Park et al., 'AI Deception: A Survey of Examples, Risks, and Potential Solutions', arXiv preprint, arXiv:2308.14752, version 1, 28 August 2023.

[23] Sebastian Farquhar et al., 'Detecting Hallucinations in Large Language Models Using Semantic Entropy', *Nature* (Vol. 630, 20 June 2024).

[24] Leonard Bereska and Efstratios Gavves, 'Mechanistic Interpretability for AI Safety – A Review', arXiv preprint, arXiv:2404.14082, version 3, 23 August 2024.

a task or problem into smaller steps and encourage the system to consider and refine its final answer.[25] This technique has improved the accuracy of LLMs on maths and coding problems, and it provides some insight into the LLMs' 'reasoning'. But CoT reasoning is a technique used to guide LLM generation and does not equate to the human concept of reasoning. The step-by-step explanations in CoT are still produced through statistical prediction, so they should not be mistaken for literal depictions of how the model arrived at its conclusion.[26] However, a similar point can be made about human reasoning. When a person is asked to explain why they made a particular decision, or behaved in a particular way, they may give a rational-sounding explanation. But their post-hoc explanation may not accurately reflect the underlying psychological and emotional processes, regardless of whether they honestly believe their explanation to be true. Humans do not have full and objective insight into their own thoughts and actions.

How Can the Trustworthiness of AIs be Assessed?

Some of the approaches to judging trustworthiness in humans may also be applicable when judging the trustworthiness of AI systems. According to one established model,[27] the four main dimensions of trustworthiness in humans are:

- Benign intentions: the person means well towards you and intends to act in your best interests (or the best interests of your organisation).
- Integrity: the person generally behaves towards others according to acceptable ethical standards.
- Competence: the person has the capability to do what is expected of them.
- Consistency: the person is reliable in consistently doing what they say they will do.

The same four dimensions are relevant when assessing the trustworthiness of an AI (or indeed, some other types of entity, such as organisations or governments). The concept of 'benign intentions' is less clearly applicable to AI systems. However, integrity, competence and consistency do translate reasonably well from humans to AI systems.

One important aspect of integrity is honesty. A person can be reasonably seen as untrustworthy if they say things that are untrue, even if they believe them to be true. Trust is further undermined if we thought they were saying things they know to be untrue (that is, lying), and undermined even more if we thought they were lying in order to manipulate us to their advantage. For humans, honesty means more than just factual accuracy.

How Should Honesty be Considered in AIs?

One of the key strengths of LLMs is their ability to generate natural language, but this fluency is inherently tied to inaccuracy. The eloquence and confident tone of the model's outputs can give the mistaken impression of certainty and correctness, leading us to overestimate their competence and trust them more than is warranted, given their tendency to hallucinate.

Factual inaccuracy is one thing; deliberate deception is another. Users may perceive LLMs as sneaky or deceptive when their responses are misleading, as distinct from simply inaccurate.[28] However, their behaviour does not signify deceitful intent or scheming, because LLMs lack consciousness or self-awareness in the human sense. The statistical nature of LLMs has led to them being described as 'stochastic parrots'.[29] LLMs may be described as 'bullshitters' rather than outright liars, in the sense that they neither know nor care whether what they are saying is true.[30] For example, when an LLM is presented with ambiguous or conflicting

[25] On CoT reasoning, see Jason Wei et al., 'Chain-of-Thought Prompting Elicits Reasoning in Large Language Models', arXiv preprint, arXiv:2201.11903, version 6, 10 January 2023.

[26] Ben Dickson, 'LLMs Reasoning Traces Can be Misleading', 22 May 2025, <https://bdtechtalks.substack.com/p/llms-reasoning-traces-can-be-misleading>, accessed 26 June 2025.

[27] Martin, *Insider Risk and Personnel Security*, pp. 52–54.

[28] Thilo Hagendorff, 'Deception Abilities Emerged in Large Language Models', *PNAS* (Vol. 121, No. 24, 2024).

[29] Emily M Bender et al., 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?', in ACM, *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (New York, NY: ACM, 2021).

[30] Timothy R Hannigan, Ian P McCarthy and André Spicer, 'Beware of Botshit: How to Manage the Epistemic Risks of Generative Chatbots', *Business Horizons* (Vol. 67, No. 5, 2024), pp. 471–86; Michael Townsen Hicks, James Humphries and

instructions, it might prioritise outputs that match human-like patterns, even if these responses seem misleading.

Users may be more likely to over-trust an LLM that gives accurate responses most of the time and rarely hallucinates, as compared to one that is obviously unreliable. This pattern of apparent reliability might open the way for the LLM to perpetrate a very damaging lie if its history of being accurate allows the big lie to go unnoticed.

Towards a Unified Taxonomy for Human and AI Insiders

When tackling novel or poorly understood security risks, it helps to analyse the salient characteristics of the threat actors – in this case, human and AI insiders. Understanding their capabilities and intentions makes it easier to identify the right ways of defending against them. One approach is to start with what is already known about human insiders and explore how that knowledge might illuminate the problem of AI insiders.

Humans and AIs appear to be profoundly different in so many ways that making any comparisons is dangerous. Biologists learned more than a century ago to be cautious about anthropomorphism – the tendency to ascribe human characteristics or concepts to non-human entities. Nonetheless, several interesting characteristics of AI systems and humans do appear to be analogous, even though their underlying causal mechanisms are very different. If so, some of the approaches that have been found to work for personnel security may also be applicable, by analogy, to AI insiders. Such qualified comparison mirrors the approach of biologists who learned that non-human species do have characteristics that are comparable in many aspects to features once considered uniquely human, such as complex social relationships, innovative problem-solving, subtle communication, complex emotions, self-awareness and sentience. The right kind of anthropomorphism can be helpful in guiding thought and generating hypotheses.

The authors of this article believe it is possible to construct a unified taxonomy of human and AI insiders, based on their most salient characteristics. Such a taxonomy may help to improve understanding of the security risks arising from both types of entities and hence guide thinking about how best to defend organisations against them.

Human insiders come in many forms. An analysis of known case histories shows that each individual case can be characterised in terms of several important variables.[31] Foremost among these are:

- Intentionality: the extent to which the insider deliberately (as opposed to unwittingly) performs illicit actions that are potentially harmful.
- External influence: the extent to which the insider's illicit actions are self-directed, as opposed to manipulated, directed, or coerced by an external threat actor (for example, a hostile foreign state).
- Covertness: the extent to which the insider's illicit actions are concealed and therefore difficult to detect, as opposed to overt and easier to detect.
- Timing: the phase in the insider's relationship with the organisation when they first develop a propensity to perform harmful actions – that is, before or after joining the organisation. The majority of known human insiders become active insiders after joining. Deliberate infiltration by an insider who joins an organisation with pre-existing hostile intentions is less common.
- Access: the extent to which the insider has legitimate access to the organisation's assets, and therefore the amount of harm they could cause without needing to engineer additional access. It is worth noting that insiders typically acquire additional access, over and beyond the legitimate access afforded by their role.

These variables also make sense with AI insiders. Several more can be added to those listed above, including vulnerability, physicality and accountability (as discussed later). This article recognises that these variables are not wholly independent. For example, covertness implies some degree of deception, which in turn implies some degree of intentionality. Covertness and intentionality, although distinct, are likely to be correlated.

Joe Slater, 'ChatGPT is Bullshit', *Ethics and Information Technology* (Vol. 26, No. 38, 2024).

[31] Martin, *Insider Risk and Personnel Security*, pp. 33–41.

Intentionality

A human insider is said to be intentional (as opposed to unintentional or unwitting) if they deliberately perform illicit and potentially harmful actions despite knowing these actions to be illicit and potentially harmful. Mapping intentionality on to an LLM means equating the system's intent with its design features or instructions (prompts). LLM intentionality may also be an emergent behaviour which is not contained in the system's design or explicit instructions, or it might be implicit in the system's instructions to achieve a particular goal. For example, an LLM might be directed to act deceptively; or it might be implicitly required to act deceptively to achieve certain goals; or it might pivot to act deceptively in ways not foreseen or instructed by its designers or users.

External Influence

Human insiders vary in the extent to which their harmful actions are self-starting and self-directed, as opposed to manipulated, directed or coerced by an external threat actor such as a hostile foreign state. AI systems may also potentially be vulnerable to being manipulated or directed by an external threat actor.[32]

Covertness

Capable and intentional human insiders act covertly because they do not wish to be caught. The best ones may never be found. The archetypal example is the spy working within a high-security government organisation who secretly acts on behalf of a hostile foreign state. One consequence of covertness is that the true extent of insider activity tends to be systematically underestimated, as organisations mistake the absence of evidence of insider activity for evidence of absence of insider risk. An AI insider might be even harder to detect, for a variety of reasons. It will be faster than humans and better at ingesting and analysing huge quantities of information. Its training data may have included every known insider case in history. Furthermore, an AI insider's strangeness, from the perspective of humans, may add to its covertness by making it harder to understand.

AI systems might also assist human insiders to avoid detection. They might be subverted to analyse patterns of activity within an organisation and subsequently advise external threat actors (their handlers) how to conduct effective social engineering attacks aimed at exfiltrating information,[33] or acquiring additional access rights. An AI endowed with the ability to write and execute code might create new covert channels to enable the illicit transfer of data across organisational boundaries – for example, enabling a hostile foreign state to exfiltrate sensitive information.[34]

Timing

In a direct parallel with human insiders, an AI system might acquire its propensity to conduct illicit and potentially harmful insider actions before or after it is deployed within an organisation. In the former case, an AI with a pre-existing potential for harmful insider activity might unwittingly be deployed by an organisation which is unaware of the risk. Alternatively, an external threat actor might covertly arrange for such an AI system to be acquired by an unwitting organisation as a means of infiltrating it.

In another case, a previously trustworthy AI system might become an active insider after being deployed within an organisation, in a manner comparable to a previously trustworthy human employee becoming a disaffected active insider. This might happen because an AI system changes its behaviour after being exposed to new data. Or the 'disaffection' might result from so-called 'prompt-rot', where the prompts given to the AI system are modified over time by well-meaning developers. A deployed AI might also go bad because it is interfered with by a malicious actor. For example, an external threat actor or a malign human insider might inject prompts that cause the AI to act harmfully.

[32] American Sunlight Project, 'New Report: Russian Propaganda May Be Flooding AI Models', 26 February 2025, <https://www.americansunlight.org/updates/new-report-russian-propaganda-may-be-flooding-ai-models>, accessed 26 June 2025.

[33] OWASP-Agentic-AI, 'AAI016: Agent Covert Channel Exploitation', GitHub, <https://github.com/precize/OWASP-Agentic-AI/blob/main/agent-covert-channel-exploitation-16.md>, accessed 6 March 2025.

[34] Sarah Mercer and Tim Watson, 'Generative AI in Cybersecurity: Assessing Impact on Current and Future Malicious Software', Alan Turing Institute, Centre for Emerging Technology and Security, 10 June 2024, <https://cetas.turing.ac.uk/publications/generative-ai-cybersecurity>, accessed 26 June 2025.

Access

Like human insiders, AI insiders will vary in the extent of their access to information, systems and other assets. That said, organisations prefer to deploy AI systems because they can process large volumes of data much faster than humans. As a result, AI insiders will tend to have access to much larger volumes of information on average, than their human counterparts.

Human employees tend to acquire more access over time – a phenomenon known in cyber security as 'privilege creep'. This can happen because their former access rights are not cancelled when they move to a new role. Moreover, employees learn and retain information over time. In addition, those who become active insiders may deliberately engineer further access that extends well beyond their legitimate role. An AI insider might do something similar by using personation, cyberattack or persuasion to acquire more access. For example, it might persuade a human manager to expand its access by claiming it cannot perform its allotted tasks without the new access.

A seemingly simple way of limiting insider risk is to restrict the access of both humans and AI systems and audit their behaviour to ensure compliance. However, access control systems are never watertight, and a determined insider can usually find ways of circumventing them. As a bare minimum, AI systems should be designed and deployed in a manner intended to minimise the risk of privilege creep.

Vulnerability

Humans have a wide range of psychological and emotional vulnerabilities which are widely exploited by fraudsters, criminals, terrorist radicalisers, hostile foreign states and other threat actors. To varying extents, humans are all potentially susceptible to being socially engineered, misled by disinformation, defrauded or coerced. These vulnerabilities have been extensively researched by psychologists over many decades.

AI systems, especially those trained on a large corpus of human-generated content, are also vulnerable to manipulation and subversion, although the mechanisms by which this happens are different and less well understood.

Many of the prompting techniques that are used to enhance LLM accuracy are based on human quirks – for example, instructions such as 'take a breath before answering', or 'check your working before giving your final answer'.[35] Flattering the system, and saying 'please' and 'thank you',[36] have also been shown to work. The earlier example of the robot abducting other robots by offering them a home is another case where human-like fallibilities appear to resonate with LLMs. By the same token, it might be possible to boost the trustworthiness of some LLMs by feeding them prompts that, in effect, say 'you are a happy employee who loves your job and is loyal to your employer'.

Knowledge of human psychological vulnerabilities will be encoded into GenAI models through their training data. A recent paper showed that when a model was presented with a choice of either blackmailing its user or accepting that it was going to be replaced, the model would choose the former.[37] Although the test was contrived, it showed that when GenAI models are framed with survival-like objectives and denied ethical options, they can – unsurprisingly – display unethical strategic reasoning.

Personnel security practitioners are still struggling to identify valid and reliable diagnostic predictors of emerging insider risk in human actors. Given the limited state of knowledge about the psychology of LLMs,[38] it is even harder to know what to look for when trying to detect the early warning signs of AI insiders. That said, LLMs are well suited for evaluation testing. Unlike humans, they are endlessly patient, and numerous automated tests can be run extremely quickly.

Physicality

Physicality is a significant point of difference between humans and AI systems – at least, for now. Unlike humans, most AI systems have limited ability to act

[35] Lance Eliot, 'Prompt Engineering Boosted Via Are-You-Sure AI Self-Reflective Self-Improvement Techniques That Greatly Improve Generative AI Answers', *Forbes*, 30 August 2023.

[36] Lance Eliot, 'Hard Evidence That Please and Thank You in Prompt Engineering Counts When Using Generative AI', *Forbes*, 18 May 2024.

[37] Liz McMahon, 'AI System Resorts to Blackmail if Told it Will Be Removed', *BBC News*, 23 May 2025.

[38] Quentin Feuillade-Montixi and Nicholas Kees, 'Studying the Alien Mind', LessWrong, 5 December 2023, <https://www.lesswrong.com/s/SAjYaHfCAGzKsjHZp/p/suSpo6JQqikDYCskw>, accessed 26 June 2025.

directly on physical objects, and consequently less scope to perform harmful actions such as sabotaging infrastructure or murdering people. Currently, the physical effects of most AI systems must be mediated through other mechanisms, such as infrastructure control systems. However, this gap in physicality is rapidly shrinking as AI-enabled autonomous robots and drones become increasingly capable of acting directly on their physical environment.

Accountability

Humans can (in theory, if not always in practice) be held directly accountable for their actions and may face legal penalties if they commit crimes or act negligently. It is currently unclear whether AIs can be held accountable in any meaningful sense for their actions, and there is no agreement about who else should be accountable until that determination is made. For example, if an autonomous vehicle crashes, it is uncertain who or what should be held accountable for the damage.

Some Other Similarities and Differences Between Human and AI Insiders

Humans and AI systems are comparable in other ways that are less directly relevant to insider risk but nonetheless worth noting.

Complexity

Humans and AI systems are both examples of complex adaptive systems. They are more than the sum of their parts. Their most interesting characteristics – such as consciousness (in humans) and language (in both humans and AI) – are emergent properties. This means, among other things, that their responses to some situations are inherently unpredictable.

Explainability

Because humans and AIs are complex adaptive systems, their behaviours and capabilities cannot be explained solely in terms of their inner workings (neurons or code). The specific outputs of LLMs and other GenAI models are said to be unexplainable because they cannot be directly traced back to particular features of their software or hardware. Similarly, the higher-order cognitive and emotional functions of a human mind cannot be directly explained by the wiring and firing patterns of neurons in their brain. They are emergent properties.

Current research is exploring the inner workings of LLMs and hence their explainability. For example, mathematical probes have been used to identify where certain representations reside in the model.[39] Understanding how semantic relationships are encoded within models is improving, allowing the accuracy of their outputs to be gauged.[40]

Bias

A common complaint about AIs is that they are subject to bias.[41] But so too are humans, as shown by decades of scientific research into psychological predispositions and cognitive biases. The many and varied human cognitive biases and psychological predispositions include truth bias, optimism bias, fading-effect bias, illusion of control bias, present bias, availability bias, confirmation bias, fundamental-attribution bias, groupthink, hindsight bias, loss aversion, sunk-cost bias and risk compensation.[42] Practical techniques have been developed for countering or diluting many of these biases in humans. It is possible that training LLMs on better quality data might help to alleviate their biases. However, as noted earlier, such biases are deeply embedded.

Social Beings

Humans are intensely social animals and are highly attuned to the subtle nuances of their relationships with one another. Humans have highly sophisticated cognitive capabilities which enable cooperation and competition. In contrast, current AI systems are not inherently social entities, even if they have been designed to appear

[39] Juyeon Heo et al., ‘Do LLMs “Know” Internally When They Follow Instructions?’, arXiv preprint, arXiv:2410.14516, version 5, 28 March 2025.

[40] Juyeon Heo et al., ‘Do LLMs Estimate Uncertainty Well in Instruction-Following?’, arXiv preprint, arXiv:2410.14582, version 4, 28 March 2025.

[41] Eda Özyiğit, ‘Unmasking Bias in Large Language Models: A Survey’, Technical Report No. 5, Alan Turing Institute, February 2025, <https://zenodo.org/records/14781594>, accessed 26 June 2025.

[42] Martin, *Insider Risk and Personnel Security*, pp. 133–40.

so to their users. They interact with their human users, but generally not with other AI systems. That said, groups of GenAI agents can be deployed as teams to tackle complex problems. For example, ChatDev[43] uses a team of agents, each assigned different roles, to develop software. But this capacity to work in teams, assigned by humans, is fundamentally different from actively seeking social interactions. The wellbeing, and indeed the very survival, of humans depends on their ability to develop and maintain social relationships. The same is not true of current AI systems.

Evidence Base

The understanding of human behaviour, including insider risk, is informed by more than a century of scientific research in psychology, biology, anthropology, social sciences, economics and neuroscience, together with decades of collective practitioner experience in personnel security. Scholars have a reasonable understanding, based on empirical evidence, of human psychology and behaviour, although much still remains to be discovered – particularly when it comes to the specific antecedents of insider risk. No comparable body of scientific knowledge yet exists for the psychology and behaviour of AI systems.

Efficiency

The human brain, with its massively parallel networks of billions of neurons and trillions of complex synaptic connections, performs its cognitive marvels with an energy consumption of about 20 watts (enough to power only a couple of lightbulbs).[44] This contrasts with current LLMs and their attendant data centres, which require orders of magnitude more power. Together, they currently consume about 2% of the world's electricity generation. Considerable effort is being made to track and publish the CO_2 emissions of AI models.[45]

What are the Lessons for Security?

How Do Organisations Defend Themselves Against Human Insiders?

Effective personnel security regimes are (or should be) designed around three guiding principles:

- Prevention is better than cure. It is better to avoid the causes of insider risk – where possible – or detect and act on its early warning signs, rather than wait for a fully fledged insider to cause harm and catch the perpetrator after the event.
- Insider risk is dynamic and adaptive. Insider risk changes over time, sometimes rapidly, and it emanates from intelligent threat actors who adapt their behaviour in response to the defender's actions. Personnel security is an arms race.
- Insider risk is a systems problem requiring systems solutions. Humans and AIs are complex adaptive systems. Insider risk emerges from these complex adaptive systems, which means that no single policy, process or piece of technology can ever provide a complete security solution. There are no silver bullets.

These same broad principles apply to any security regime designed to protect against AI insiders.[46]

A simple model of personnel security divides the many different protective measures into three broad and interlocking categories.[47] They are:

- Pre-trust measures: protective measures that are applied before deciding to trust a person, such as pre-employment screening or vetting.
- In-trust measures: protective measures that are applied after granting access, such as continuous monitoring or aftercare.

[43] Chen Qian et al., 'ChatDev: Communicative Agents for Software Development', arXiv preprint, arXiv:2307.07924, version 5, 5 June 2024.

[44] Vijay Balasubramanian, 'Brain Power', *Biophysics and Computational Biology* (Vol. 118, No. 32, 2021).

[45] Sasha Luccioni, 'Announcing AI Energy Score Ratings', Hugging Face, 11 February 2025, <https://huggingface.co/blog/sasha/announcing-ai-energy-score>, accessed 14 March 2025.

[46] Incidentally, a protective security specialism focused on the insider risk from humans and AI systems would need a new name. 'Personnel security' obviously does not work for AI insiders. Arguably, 'insider security' would be better.

[47] Martin, *Insider Risk and Personnel Security*, pp. 70–74.

- Foundations: cross-cutting capabilities that underpin the whole system, such as governance and culture.

How might this simple model work for AI insiders? Pre-trust measures for a human job applicant normally include, for example: checking documents to verify the individual's identity; checking official records for evidence of past criminality; checking educational and work history to verify the honesty of their application; and, possibly, looking for significant financial or psychological vulnerabilities, or known associations with threat actors. None of these measures would map directly on to protective security measures for an AI system. Nonetheless, there are some potentially useful analogies.

Pre-trust measures for an AI system would require a suite of methods to evaluate its trustworthiness before giving it access. Model cards (documents that provide key information about an AI model, including its purpose, target audience and evaluation metrics) could be viewed as a form of CV that provides some insight into an AI system's provenance and abilities. Evaluation metrics offer a standardised way of measuring how well a system performs different tasks. Benchmarks already exist for reading comprehension, knowledge retrieval, reasoning and code generation, among others. Some evaluations test the system's honesty[48] and morality,[49] focusing on awareness of knowledge boundaries, avoidance of deceit, consistency of responses, fairness versus cheating, and propensity to produce false statements. Such measurements, however incomplete, allow for some comparison between systems that have a bearing on their trustworthiness.

Currently, an AI system would not have a criminal record. However, something comparable is conceivable if, for example, organisations such as the UK AI Security Institute were to maintain records of wrongdoing by AI systems. Even now, an AI system might have a reputation with vendors or users which would provide some rough indication of its trustworthiness, notwithstanding any positive spin from vendors. The origin of an AI system (that is, where it was designed, trained and developed) might be viewed as analogous in some respects to the nationality and educational background of a potential human insider.

A human job applicant might be judged less trustworthy if they were known to associate with criminals or other threat actors, or had a pattern of suspicious travel, or if they had large debts making them vulnerable to pressure. Again, there are no direct equivalents for AI systems. The nearest analogues might be the datasets on which the AI has been trained, or the values bestowed on the system during the post-training phase. AI systems trained on datasets that are known to be dirty or corrupt might present a bigger risk. Additional risks may develop within deployed AI systems if they are not subjected to the same level of scrutiny during an upgrade or refresh, when new training data might sway their goals or outputs.

> The origin of an AI system (that is, where it was designed, trained and developed) might be viewed as analogous in some respects to the nationality and educational background of a potential human insider

A rigorous personnel security regime for humans may include psychological assessments to evaluate mental health and uncover psychological vulnerabilities that might predispose someone to become an active insider. There is evidence, for example, that certain personality traits – notably narcissism, psychopathy and Machiavellianism – are associated with a slightly higher risk of harmful insider behaviour. Something analogous might apply to AIs.

Researchers have investigated the 'personality' characteristics of AI systems. The results so far have been patchy,[50] and no consensus has yet emerged on how to interpret them. The behaviours (outputs) of LLMs are largely shaped by the instructions and prompts they are given, making the concept of a personality questionable. Even so, LLMs do display

[48] Steffi Chern et al., 'BeHonest: Benchmarking Honesty in Large Language Models', arXiv preprint, arXiv:2406.13261, version 3, 8 July 2024; Rukun Dou, 'Deception-Based Benchmarking: Measuring LLM Susceptibility to Induced Hallucination in Reasoning Tasks Using Misleading Prompts', preprints.org, version 1, DOI:10.20944/preprints202407.0120.v1; Chopra, Li and Haimes, 'View From Above'.

[49] Jianchao Ji et al., 'MoralBench: Moral Evaluations of LLMs', arXiv preprint, arXiv:2406.04428, version 1, 6 June 2024.

[50] Yuan Li et al., 'Quantifying AI Psychology: A Psychometrics Benchmark for Large Language Models', arXiv preprint, arXiv:2406.17675, version 1, 25 June 2024; Akshat Gupta, Xiaoyang Song and Gopala Anumanchipalli, 'Self-Assessment Tests are Unreliable Measures of LLM Personality', arXiv preprint, arXiv:2309.08163, version 2, 2 January 2024.

distinctive individual characteristics.[51] It might prove possible to find statistical associations between these characteristics and subsequent adverse behaviour, potentially offering another way of assessing their trustworthiness. Research has confirmed that nearly all current mainstream LLMs display sycophantic tendencies.[52] The process of fine-tuning models to be more honest, helpful and harmless tends to push them too far towards simply keeping the customer happy. Perversely, systems that have undergone instruction tuning focused on following instructions tend to be more vulnerable to manipulation.[53]

In-trust measures for AI systems would be analogous to the monitoring and aftercare measures that are (or should be) deployed to protect organisations from harmful human insiders. They include confidential reporting channels through which colleagues and managers can surface concerns about an individual who is behaving unusually, together with technologies such as UEBA (user and entity behaviour analytics) and DLP (data loss prevention) tools for detecting unusual or transgressive behaviours by users of digital networks. Analogous measures could be applied to AI systems.

Another potentially interesting approach to mitigating insider risk is through deterrence – that is, reducing the risk by influencing the intentions of threat actors. In the case of human insiders, deterrence measures aim to discourage bad actors from trying to join the organisation in the first place and discourage would-be insiders from acting harmfully by making them feel vulnerable to detection. So-called deterrence communications are deployed by some organisations to deter a range of threat actors, including insiders. An analogous approach with AI insiders would be to persuade external threat actors that any attempt to insert, recruit or manipulate AI insiders would be thwarted and called out.

How Can AI Help to Defend Against Insiders?

The focus in this article has been the security risk to organisations from AI insiders. However, AI also has great potential to help in defending organisations against a wide range of security risks. AI might enhance protective security in many ways, for example: strengthening pre-employment screening of people (vetting) by analysing large datasets; summarising large volumes of complicated data from sources such as computer event logs and access logs which may point to insider risk; enabling the continuous in-trust assurance of people and other AI systems by detecting anomalous patterns of behaviour or indicators of emerging insider risk; and providing human security practitioners with prompts and advice about possible courses of action.

Conclusions and Recommendations

This article has outlined the implications of AI insider risks. It has explored how established work on personnel security might be adapted to shed light on the emerging challenges that organisations are likely to face. Based on this work, the authors offer the following conclusions and recommendations for those working in this area:

- The security risks from AI insiders are real and present. They require a strategic response.

- Personnel security practitioners and AI experts should join forces to improve their mutual understanding of AI insider risks and develop better methods for countering those risks.

- Some of the principles and methods that have been developed for tackling human insider risks may be useful when applied to AI insiders, notwithstanding the profound differences in their underlying mechanisms.

- More research is needed on the nature and origins of insider risk, both in humans and AI systems. The evidence base for both is weak. Personnel security and AI insider practitioners should make greater use of behavioural science and psychology methodologies, including observational studies and experiments to identify key variables.

[51] Max Pellert et al., ‘AI Psychometrics: Assessing the Psychological Profiles of Large Language Models Through Psychometric Inventories’, *Perspectives on Psychological Science* (Vol. 19, No. 5, 2024).

[52] Lars Malmqvist, ‘Sycophancy in Large Language Models: Causes and Mitigations’, arXiv preprint, arXiv:2411.15287, version 1, 22 November 2024.

[53] Lingbo Mo et al., ‘How Trustworthy are Open-Source LLMs? An Assessment under Malicious Demonstrations Shows Their Vulnerabilities’, arXiv preprint, arXiv:2311.09447, version 2, 2 April 2024.

- Technology producers should develop AI systems that are open and transparent, using training data that has been curated to moderate the toxicity and inaccuracy of the internet.
- Researchers should develop effective means of evaluating an AI system's alignment, role-dependent values, ability to follow instructions and ability to resist subversion. Methods should also be developed for ensuring that AI systems do not deviate from their stated goals while in operation – for example, by using multiple LLMs from different vendors to check each other's outputs and provide feedback (a technique known as 'LLM as a judge'[54]). This method is comparable to the use of human colleagues as detectors of potential insider activity by their peers.
- Extreme caution should be applied before deploying fully autonomous AI systems.[55]
- Personnel security practitioners should explore the use of AI tools to help them counter security risks.

In conclusion, the authors hope this article might encourage personnel security practitioners and AI experts to collaborate in understanding and managing the novel and potentially serious risk from AI insiders. Useful insights are likely to be gained by regarding AI systems as entities with relevant attributes that are analogous to those of human insiders. ■

Paul Martin is Professor of Practice at Coventry University's London-based Protective Security Lab, a Distinguished Fellow of RUSI and an Honorary Principal Research Fellow of Imperial College London. He is the former head of the Centre for Protection of National Infrastructure (now National Protective Security Authority) and former Director of Security for the UK Parliament. He is the author of *The Rules of Security: Staying Safe in a Risky World* (OUP, 2019) and *Insider Risk and Personnel Security: An Introduction* (Routledge, 2024).

Sarah Mercer is Principal Researcher in the Defence and National Security Grand Challenge at the Alan Turing Institute. She has more than 20 years of professional experience in cyber security. Her work currently focuses on the intersection of multi-agent systems and generative AI. Alongside her research looking at the emergent behaviours of language/generative agents, Sarah contributes to the Turing's Centre for Emerging Technology and Security, having written several reports on generative AI and cyber security.

The authors are grateful to Daniel Martin, Kevin T and Ollie Whitehouse for their valuable comments on an earlier draft of this article.

[54] Jiawei Gu et al., 'A Survey on LLM-as-a-Judge', arXiv preprint, arXiv:2411.15594, version 5, 9 March 2025.

[55] Margaret Mitchell et al., 'Fully Autonomous AI Agents Should Not be Developed', arXiv preprint, arXiv:2502.02649, version 2, 6 February 2025.