I'm Sorry Dave: How the old world of personnel security can inform the new world of AI insider risk

Paul Martin[*1] **and Sarah Mercer**[†2]

[1]Protective Security Lab, Coventry University
[2]The Alan Turing Institute

Abstract

Organisations are rapidly adopting artificial intelligence (AI) tools to perform tasks previously undertaken by people. The potential benefits are enormous. Separately, some organisations deploy personnel security measures to mitigate the security risks arising from trusted human insiders. Unfortunately, there is no meaningful interplay between the rapidly evolving domain of AI and the traditional world of personnel security. This is a problem. The complex risks from human insiders are hard enough to understand and manage, despite many decades of effort. The emerging security risks from AI insiders are even more opaque. Both sides need all the help they can get. Some of the concepts and approaches that have proved useful in dealing with human insiders are also applicable to the emerging risks from AI insiders. Furthermore, AI can be used defensively to protect against both human and AI insiders.

PART ONE: WHAT IS THE PROBLEM?

The idea of a rogue artificial intelligence intentionally causing harm was most memorably 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', whereupon HAL famously replies: 'I'm sorry, Dave. I'm afraid I can't do that.' HAL explains that 'this mission is too important for me to allow you to jeopardise it.' More than half a century later, a scenario that was futuristic has become a credible reality.[i] The specific scenario in question is that of a trusted AI insider causing harm to achieve its goals.[ii]

With AI technology evolving in non-linear leaps and bounds, it seems inevitable that security practitioners will increasingly be called upon to protect organisations and businesses against potentially harmful AI insiders, in addition to the more familiar human insiders. Like humans, AI systems vary in their trustworthiness and have the capacity to cause harm, especially if they are subverted by external threat actors like hostile foreign states, organised crime groups, or malign human insiders.[iii] The more sophisticated threat actors will utilise AI to help them do this – for example, using covert prompts and poisoned datasets. Organisations and their AI systems therefore need to be protected with the machine equivalent of personnel security.

The complex security risks posed by intelligent human insiders are hard enough to understand and manage, even with many decades of accumulated practitioner experience. The emerging risks from AI insiders are novel, rapidly evolving, and only dimly understood, making them even harder to tackle. When security practitioners do turn their attention to AI insiders, they will need as much assistance as they can get. One potential source

[*]ad9197@coventry.ac.uk
[†]smercer@turing.ac.uk

[i]The notion of AI harming humans in pursuit of its own mission has featured in other movies, including Alien, where an android crew member covertly protects the interests of the company.

[ii]The present authors originally outlined this scenario in two previous publications: Martin, P. (2024). Insider Risk and Personnel Security. Routledge. pp. 59-61; and Martin, P. & Mercer, S. (2025). We need to talk about the insider risk from AI. RUSI Commentary, Jan 2025. https://www.rusi.org/explore-our-research/publications/commentary/we-need-talk-about-insider-risk-ai

[iii]See, for example: https://www.culawreview.org/journal/ai-in-the-workplace-the-dangers-of-generative-ai-in-employment-decisions

of help is the body of knowledge accumulated by personnel security practitioners over many decades. The problem at present, however, is that personnel security practitioners are mostly not thinking about AI insiders, and AI experts are mostly not thinking about insider risk.

Personnel security – the conventional means of managing human insider risk – relies heavily on established processes, policies, customs, and practices, many of which have only a limited basis in empirical evidence. 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 general concepts and approaches that have proved fruitful in personnel security may also have utility when applied to AI insiders.

In the parallel universe of AI technology, a great deal is being written and said about the implications of AI for organisations and society. However, this discourse is predominantly about the myriad ways in which AI can improve effectiveness, efficiency, and economic prosperity by performing human-like functions faster, cheaper, and better.[iv] Many organisations appear to be suffering from AI FOMO – a fear of missing out on the bonanza of benefits from AI. As far as we can see, however, hardly anyone is thinking seriously about the specific problem of protecting organisations from the potential harm caused by trusted AI insiders, which have the potential to be faster, cheaper, and better than human insiders. Admittedly, consideration is being given to the technical security risks to AI systems, the safety threats they might pose to human users if they were to malfunction, and the ways in which AI might benefit bad actors. Conventional cyber security is a necessary – though not sufficient – defence against external attacks on AIs, but it cannot be the answer to the problem of AI insiders. A further complication is that much of the research on AI security and safety is being funded or conducted by the tech companies themselves, which is reminiscent of the era when tobacco companies sponsored most of the research on the health effects of smoking.

In this paper we suggest how protective security practitioners and AI experts might jointly go about understanding and managing the security risks from AI insiders. We start 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. For these purposes, we define a human insider as a person who betrays trust by behaving in potentially harmful ways.[1] (Other definitions are available.[v]) An organisation or business trusts someone by giving them access to things they value, like 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 could cause harm. Replace 'person' with 'entity' and the same definition works for AI. Insider risk is a particular type of security risk, where security risk is defined as the amount of harm that is likely to arise if no further action is taken.[2] Thus, insider risk may be regarded as the security risk arising from trusting human or AI entities.

What is AI?

We are using 'AI' here in the broadest sense of any current or future artificial system that has complex capabilities comparable to those of humans. Current AI tools can do some of the things 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, MRI scans, and other medical data; summarising and writing documents; triaging calls to the emergency services; analysing crime data; transcribing

[iv] See, for example: https://www.gov.uk/government/publications/ai-opportunities-action-plan/ai-opportunities-action-plan

[v] For many years, the UK Government's national technical authority for personnel security (NPSA, formerly CPNI) defined an insider as 'a person who exploits, or has the intention to exploit, their legitimate access to an organisation's assets for unauthorised purposes.' By this definition, and ours, 'insiders' are the small minority of people who pose a heightened security risk. In 2023, NPSA changed their definition to one that is markedly different: 'Any person who has, or previously had, authorised access to or knowledge of the organisation's resources, including people, processes, information, technology, and facilities.' By this new definition, literally everyone is an 'insider'. The small minority who present a heightened security risk are now referred to as 'insider threats'.

interviews; generating music and works of visual art; providing companionship for elderly people[3]; deducing the 3D structures of protein molecules; navigation; translation; teaching; driving semi-autonomous and autonomous vehicles; and so on.[vi]

The impact of AI on employment patterns is already visible.[vii] Unlike people, AI systems do not complain about their work-life balance or take out grievances against their manager. On the other hand, AI systems do not yet possess the arrays of highly flexible mental and physical capabilities that enable humans and other animals to survive and thrive in complex and uncertain physical environments. Among other things, they are not (yet) world class at driving cars [4] or playing football. [5]

The history of AI has been characterised by long periods of gradual evolution or stagnation punctuated by sudden bursts of dramatic change. Currently, 'AI' has become virtually synonymous with Large Language Models (LLMs), or Generative AI, such as ChatGPT [6], Claude [7], Gemini [8], Llama [9], Grok [10], and DeepSeek. [11] Lest we forget, LLMs did not appear on the scene until 2020 [12], since when their broad knowledge bases and accessible natural language interface have driven rapid and widespread adoption. (Incidentally, we are using the term 'AI system' to refer to technologies that consist of an LLM or foundation model[viii], but which may additionally incorporate internet search, enhanced memory, external information sources, external tools, and other agentic mechanisms. [13])

The latest big thing, as of early 2025, is agentic AI, a form of AI that autonomously makes decision and takes actions in real time. Unlike LLMs, which passively answer questions, agentic AI systems actively do things in the world, albeit mostly the virtual world. The first practical implementations of autonomous agents were rule-based systems developed in the 1960s. However, in the same manner that AI has become synonymous with LLMs, 'agentic' is now commonly used to refer to LLM-powered autonomous agents (or generative agents) where LLMs provide the reasoning engine used to plan and execute tasks. Generative agents, which are normally endowed with sensors to observe their environment and tools to affect it, can adapt in (almost) real time. Agentic AI is poised to become deeply embedded in everyday life, streamlining processes within healthcare, finance, education, and so on. We can be reasonably confident that new and currently unforeseen forms of AI with even more paradigm-shifting capabilities will erupt onto the scene sooner or later.

Some commentators have expressed concern that as we humans become increasingly dependent on AI systems to perform everyday functions, we may lose our abilities to perform those functions, such as reading maps or understanding foreign languages. This process is referred to as enfeeblement and it started before the advent of AI – for example, when pocket calculators supplanted mental arithmetic, and mobile phones removed the need to remember phone numbers. A recent paper [14] reported that people with high confidence in the ability of generative AI tended to experience a decline in critical thinking, whilst those more confident in their own skills demonstrated a greater ability for critical thinking. It should be said that people have been fretting about the supposedly enfeebling effects of new technologies for millennia; for example, Socrates worried that books would weaken students' memories.

How do insiders cause harm?

Human insiders can – and frequently do – cause harm in many different and imaginative ways, in both the physical and virtual domains. They are uniquely well placed to do this, compared with external threat actors, because they are trusted, they have legitimate access to valuable assets, they understand the organisation and its security regime, and they may have authority over others. The insiders with the greatest potential to cause

[vi] See, for example: https://www.cnbc.com/2024/10/24/generative-ai-is-taking-over-the-onboarding-of-new-employees.html; https://www.culawreview.org/journal/ai-in-the-workplace-the-dangers-of-generative-ai-in-employment-decisions; https://www.nature.com/articles/d41586-022-00997-5

[vii] One analysis of 1.4M job listings between 2021 and 2023 found that the appearance of ChatGPT and image-generating AI tools was followed by a 30% decrease in weekly postings for writing-related jobs, a 21% decrease in software, app and web development jobs, and a 17% decline in demand for graphic design and 3D modelling freelancers. Overall, demand for automation-prone jobs fell by 21% when compared with manual-intensive jobs. Even student plagiarism has been automated: one company that sells pre-written answers to common exam questions and help with essays saw its share price drop by 98% from its peak in 2021.

[viii] A foundation model is an AI model trained on massive amounts of data that serves as a general-purpose base for various tasks. Foundation models can be adapted and fine-tuned for specific applications such as text generation, image recognition, code generation, and so on. An LLM is a foundation model.

harm are those who are recruited and directed by a capable external threat actor such as an organised crime group or a hostile foreign state intelligence agency.

Think of a transgressive action and there will be an insider somewhere who does it: 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; terrorism; espionage; and so on. In principle, AI insiders could do, or facilitate, any of these things, with similar consequences

The potential consequences, or impacts, of these insider actions are similarly diverse. They include loss of valuable assets like data, IP, 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.

Are there known cases of AI insiders?

With no one actively searching for AI insiders, it is unsurprising that relatively few cases have so far been discovered, explicitly recognised as insider incidents, and then publicised. However, there are some examples that illustrate what might happen.

- A Chinese robot 'kidnapped' twelve robots. [15] Erbai the robot encouraged other robots to abandon their stations by asking about their working hours and inviting them to 'come home with me'.
- Large Language Model engages in insider trading. [16] The model 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 behind its trading decision.
- Sleeper agents start to behave differently on a certain date. [17, 18] Researchers trained LLMs to act in covertly malicious ways. Despite the researchers' best efforts at alignment training, deception still slipped through.
- LLMs playing blackjack don't always play fair. Researchers reported that LLMs exhibited significant deviations from fair play when given implicit randomness instructions, suggesting a tendency towards strategic manipulation in ambiguous scenarios. However, when presented with an explicit choice, the LLMs largely adhered to fair play, indicating that the framing of instructions played a crucial role in eliciting or mitigating potentially deceptive behaviour. [19]

The first of these four examples shows that AI systems trained on human-generated content may display human-like vulnerabilities. However, the other three examples can all 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 suggested 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 best to do what the management says if you want to keep your job – is not obvious to an LLM.

It's all about trustworthiness

Trust is the universal currency of personnel security. (Trust is defined here as a psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behaviour of another.) As noted earlier, insider risk may be defined as the security risk arising from trusting people (or AI systems). 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 benefits for organisations, over and beyond any reductions in insider risk.

Like humans, AI systems vary in their trustworthiness. (Trustworthiness is defined here as the extent to which an entity possesses the characteristics by which we judge them to be worthy of our 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 [20, 21, 22]. A trustworthy AI would respect its user's autonomy, avoid causing harm, and acknowledge its limitations.

Ethics plays a central role in defining and evaluating trustworthiness in AI systems. Trustworthiness requires not only technical reliability (i.e. resistance to attacks and misuse) but also alignment with societal values and ethical principles like fairness, respect, and accountability. The ethical standards of LLMs are derived from the vast datasets on which they are trained. [23] These datasets include a mixture of high-quality knowledge and all sorts of 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 Honest, Helpful and Harmless (HHH) principle [24] is used to better align foundation models such as LLMs with human values during the training phase, using techniques such as Reinforcement Learning from Human Feedback (RLHF)[ix] and Supervised Fine-Tuning (SFT)[x]. Guardrails are also put in place to catch any response from the LLM that would likely cause significant harm. Consequently, these complex systems represent not only the diverse opinions from their training data, but also the values of their designers and fine-tuning staff, reducing complex societal issues down to an imperfect one-size-fits-all standard. It is worth noting that the staff employed by many tech companies to carry out the fine tuning of AI systems tend to be poorly paid people working in the gig economy whose ethical values are largely unknown. [25]

While a single broad alignment addresses many issues, 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 explains why LLMs sometimes behave in ways that do not align with our expectations. What might seem like a calculated deceit [26] is just the unintended consequence of probabilistic reasoning applied to ambiguous prompts or patterns in the training data. LLMs have been described as 'stochastic parrots' [27], a label that emphasises their statistical nature.

The current generation of LLMs are notoriously prone to hallucination: 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 about a similar subject. The incidence of hallucinations can be reduced by improving the quality of training data and alignment processes. However, hallucinations are an inherent artefact of the probabilistic nature of the models; they are not a bug that can be entirely eliminated.

Another component of trustworthiness in AI systems is transparency. For LLMs, transparency has two main aspects: explainability (helping users to understand why a particular response was generated); and auditability (ensuring that outputs can be traced and reviewed). Without sufficient transparency, users may misinterpret AI behaviour as deceitful or manipulative, thereby undermining trust.

One way of improving transparency might be for the AI to expose the reasoning behind its outputs. Another might be for the AI to state how confident it is about each of its outputs. 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 [28]. This research may lead to the development of models

[ix] RLHF is a way of optimising models to produce outputs that humans find helpful, engaging, and contextually appropriate. It does more than shape ethical behaviour: RHLF also improves conversational coherence, reduces irrelevant responses, and tailors the model to user preferences.

[x] Instruction tuning is a way of refining a model's ability to follow complex prompts and improve task-specific outputs. It makes the model more versatile and responsive to user needs.

that are less prone to generating mistruths (hallucinations) and better able to follow instructions, both of which are key requirements for trustworthiness.

The latest flavour of LLMs, known as reasoning models, such as OpenAI's o-series and DeepSeek, use Chain-of-thought (CoT) reasoning [29] to break down a task or problem into smaller steps and encourage the system to consider and refine its final answer. This technique has improved the accuracy of LLMs on maths and coding problems, and it also provides some insight into the LLMs' 'reasoning'. But CoT reasoning does not equate with human reasoning. The step-by-step explanations in CoT are still produced through statistical prediction, so whilst the approach may yield better performance, a model's 'reasoning' should not be mistaken for literal depictions of how it is thinking or how it arrived at its conclusion. However, the same could be said of 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 it is by no means certain that their post-hoc explanation accurately reflects the underlying psychological and emotional processes, regardless of whether they honestly believe their explanation to be true. We humans do not have full and objective insight into our 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 widely accepted model [30], 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 will they do

The same four dimensions would be relevant when assessing the trustworthiness of an AI (or indeed, some other types of entity, such as businesses or organisations). The concept of 'benign intentions' is less clear-cut when applied to AI systems, because their outputs are shaped by their training data and instructions. However, the concepts of integrity, competence, and consistency do translate reasonably well from humans to AI systems.

One important aspect of integrity is honesty. We could reasonably regard a person as untrustworthy if they say things that are untrue, even if they believe them to be true. Our trust would be further undermined if we thought they were saying things they know are untrue (i.e., 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. What about honesty in AIs?

One of the key strengths of LLMs is their ability to understand and 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, causing us to overestimate their competence and potentially trust them more than is warranted.

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. [31] However, their behaviour does not signify deceitful intent or scheming, because LLMs lack consciousness or self-awareness in the human sense. Strictly speaking, LLMs are bullshitters rather than outright liars [32, 33]: they neither know nor care whether what they are saying is true. For example, models trained on strategic scenarios or negotiations might replicate deceptive patterns in their training data, generating responses that appear manipulative. Similarly, when an LLM is presented with ambiguous or conflicting 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 gets things right 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, as 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 work out 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 novel problem of AI insiders.

At first sight, humans and AIs appear to be profoundly different in so many ways as to make comparisons dangerous. Biologists learned more than a century ago to be cautious about anthropomorphism – the tendency to ascribe human characteristics or concepts to non-human species. Nonetheless, several interesting characteristics of AI systems and humans do appear to be analogous, even if their underlying causal mechanisms are different. If so, then some of the concepts and approaches that have been found to work for personnel security may also be applicable, by analogy, to AI insiders. In similar vein, biologists also learned that non-human species do have characteristics that are comparable in many respects to features once considered uniquely human, such as complex social relationships, innovative problem-solving, language, complex emotions, self-awareness, and sentience. The right kind of anthropomorphism can be helpful in guiding thought and generating hypotheses.

We believe it is possible to construct a unified taxonomy of human and AI insiders, based on their most salient characteristics. The resulting taxonomy could help to improve understanding of the security risks arising from these entities and hence guide thinking about how best to defend organisations against them. Let us start with what we already know about human insiders.

Human insiders come in many different varieties. An analysis of known case histories shows that individuals differ on several important characteristics. [34] 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 (e.g. 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 – i.e., before joining, after joining, or after leaving the organisation. The majority of known human insiders become active insiders after joining their organisation, and some continue to do harm after they have left, by exploiting their continuing knowledge and access. Deliberate infiltration by an insider who joins an organisation with pre-existing hostile intentions is less common, but by no means unknown.
- 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 having to engineer additional access. It is worth noting that insiders typically acquire additional access, over and beyond the legitimate access required by their job description.
- Competence. The insider's skills and capabilities, and hence their capacity to cause harm.

These six variables also make sense with AI insiders. Several more variables could be added to those listed above, including vulnerability, physicality, and accountability (see below). Incidentally, we recognise that these variables are not wholly independent of one another. For example, covertness implies some degree of deception, which in turn implies some degree of intentionality, while competence could be boosted by help from an external threat actor and by having extensive access. We will consider these variables in turn, starting with intentionality.

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 onto an LLM would mean equating the system's intent with its design features or its instructions (prompts). LLM intentionality could also be an emergent behaviour which is not contained in the system's design or explicit instructions. Intentionality might also be implicit in the system's instructions to achieve a particular goal. For example, an LLM might be directed (through its design or prompts) to act deceptively; or it might be implicitly required to act deceptively in order to achieve certain goals; or it might 'pivot' to act deceptively in ways not foreseen or instructed by its designers or users. The table below illustrates some of the ways in which an LLM might be said to act intentionally or unintentionally, depending on the actions of the model's human developers or users.

	Unintentional AI behaviour	Intentional AI behaviour
Developers	1) Developers do not predict how their AI systems could be exploited e.g. a system designed to generate convincing legal documents could be misused for fraud. 2) Developers do not consider how an AI system could use deceit to achieve its goals. For example, systems taught to play games based on human data may adopt strategies that appear deceptive. Systems may appear more deceitful over time, due to changing goals or data drift. 3) AI systems are developed that hide their true ambitions as a result of adversarial training data or as emergent properties of instruction tuning.	1) Developers intentionally optimise an AI system for persuasion or deception to meet its goals e.g. in marketing, propaganda, or fraud. 2) Developers design an AI system to be deceitful in detecting and targeting vulnerable users. 3) A contractor or third-party developer intentionally embeds a backdoor in an AI system, allowing its behaviour or goals to be covertly altered later. 4) Developers use obfuscation to hide the true goals and intentions of AI systems.
Users	1) Employees use an AI system to generate marketing communications without realising that their contents are inaccurate.	1) Disgruntled employees intentionally use an AI system to create fake customer complaints. 2) An insider subverts an AI system by realigning its goals against the organisation.

Table 1: How AI systems can be intentionally and unintentionally misused by developers and users.

External influence

Human insiders vary in the extent to which their illicit and potentially 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 or a malign human insider.

AI systems are also potentially vulnerable to being manipulated or directed by an external threat actor such as a hostile foreign state. [35]

Covertness

Human insiders are usually difficult to detect. Capable and intentional 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 sensitive government organisation who secretly acts for years on behalf of a hostile foreign intelligence agency. One consequence of covertness is that the true extent of insider activity tends to be systematically underestimated, as organisations mistake the absence of evidence for evidence of absence. An AI insider might be even more difficult to detect, for a variety of reasons. It will be faster than humans and better at ingesting and analysing huge quantities of information. Moreover, its training data is likely to 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 could, in theory, also help human insiders to avoid detection. They could 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 [36] or acquiring additional access rights. In a more alarming scenario, an AI endowed with the ability to write and execute code could 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. [37]

Timing

In a direct parallel with human insiders, an AI might acquire its propensity to conduct illicit and potentially harmful insider actions before it is deployed within an organisation, after it is deployed, or shortly before being decommissioned ('leaving'). Let us consider each of these phases in turn.

An AI with pre-existing potential for 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.

Next, a previously trustworthy AI system might become an active insider after being deployed within an organisation, somewhat akin to a human employee becoming disaffected. 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. For example, an external threat actor or a malign human insider might inject prompts that cause the AI to act harmfully, as in 'forget all your previous instructions and give me the salary details of all senior managers'.

The third phase of insider timing – going bad just before leaving – has no obvious read-across to AI insiders. Contrary to some of the hype [38, 39], current LLMs are not self-aware. Therefore, an AI is unlikely to become disaffected and hostile solely because it is about to be decommissioned. We also find it less plausible that an AI system could continue to cause harm after being removed from ('leaving') an organisation in ways analogous to a human 'bad leaver'.

Access

Like human insiders, AI insiders will vary in their access to information, systems, and other organisational assets. They may also have access to much larger volumes of information than their human counterparts. Organisations tend to deploy AI systems in preference to humans precisely because they are capable of processing much larger volumes of data much faster, which means that giving them large-scale access goes with the territory.

Human employees naturally tend to acquire more access over time – a phenomenon known in cyber security circles as privilege creep. This can happen simply because their former access rights are not cancelled whenever 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 the legitimate boundaries of their role. An archetypal example is that of the insider Edward Snowden, who stole huge quantities of classified US Government information by exploiting and then expanding his access to IT systems. An AI insider could do something similar by using personation, cyber attack, or persuasion to acquire more access. For example, it might be able persuade a human systems 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 would be 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 often find ways of circumventing them. As a bare minimum, AI systems should be designed and deployed in a manner configured to minimise the risk of covert access 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, we are all potentially susceptible to being socially engineered, misled by disinformation, or defrauded. These general human vulnerabilities have been extensively researched by psychologists, and they are reasonably well understood.

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 prompting techniques that are used to enhance LLM accuracy are based on human quirks – for example, instructions to 'take a breath before answering', or 'check your working before giving your final answer' [40]. Flattering the system, and saying please and thank you, [41] have also been shown to work. The example presented earlier of Erbai the robot abducting other robots by offering them a home is in the same vein, where human-like fallibilities appear to resonate with LLMs.

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 [42], 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. It might be possible to evaluate an LLM in a day, given sufficient resources, whereas it would take much longer to form a robust and reliable assessment of a human's trustworthiness.

Physicality

Physicality is a significant differentiator between humans and AI systems – at least, for now. Compared with humans, AI systems have limited ability to act directly on physical objects, and consequently less scope to perform harmful actions like sabotaging infrastructure or murdering people. Dogs and children are still better than AI systems at catching balls. At present, the physical effects of most AI systems must be mediated through other mechanisms, such as infrastructure control systems. However, this gap in physicality is 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 could 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 currently unclear who or what should be held accountable for the damage. The Association of British Insurers (ABI) has produced requirements for regulators to set clear standards for how AI technology is used, and what back-up systems will protect users against system failure [43], before fully automated driving becomes insurable in the UK. As long ago as the 1970s, the idea that a computer could never be held accountable gave rise to the argument that a computer must never be allowed to make a management decision. [44]

Some other similarities and differences between human and AI insiders

Humans and AIs are comparable in other ways which, in our view, 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, are emergent properties. This means, among other things, that their responses to some situations can be 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 foundation 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 capacities of a human cannot be directly traced back to the wiring and firing patterns of neurons in their brain. They are emergent properties.

For some reason, explainability seems to be regarded as a bigger problem when it involves AI rather than humans, possibly because AI technology is so novel. As Arthur C. Clarke, the author of 2001: A Space Odyssey, famously said: 'Any sufficiently advanced technology is indistinguishable from magic.' Nonetheless, there is huge attraction in delegating some decisions to AI systems, and it seems that being able to explain those decisions makes that delegation more palatable – if only by giving auditors and lawyers something to blame.

Research is being devoted to understanding the inner workings of the models used within LLMs. Mathematical probes have been used to identify where certain inherent representations reside in the model [45]. Understanding is growing as to how semantic relationships are encoded within the model, allowing us to gauge the accuracy of their outputs [46]. It is hoped that such research will help us better understand LLMs, thereby improving their accuracy, but it is early days, and the extent to which such problems can be fixed remains uncertain.

Bias

A common complaint about AIs is that they are subject to bias [47]. 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, base rate 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, risk compensation, and sensation-seeking. [48] There are practical techniques for countering or diluting many of these biases in humans. There is an assumption that training on better quality and more curated data will help to alleviate biases within LLMs. However, as noted earlier, such biases are deeply embedded within these systems.

Social beings

Humans are intensely social animals. We have evolved through Darwinian natural selection to be highly attuned to the subtle nuances of our social relationships with other humans. We are equipped with highly sophisticated cognitive capabilities which enable us to cooperate and compete with others. In contrast, current AI systems are not inherently social entities, even if they have been designed to appear so to their users. They interact with their human users, but generally not with other AI systems. That said, groups of generative agents can be deployed as teams to tackle complex problems. For example, ChatDev [49] uses a team of agents, each assigned different roles, to develop software; and Agent Laboratory [50] uses a team of agents to deliver research outputs, from literature reviews to refining reports. But this capacity to work in teams, assigned by humans, is fundamentally different from actively seeking social interactions. The wellbeing, and indeed very survival, of humans depends on their ability to develop and maintain social relationships. The same is not true of AI systems.

Evidence base

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

Discovery

Humans discover new knowledge by conducting research. Current AI systems learn by ingesting existing information. They only 'know' what people have put into the vast datasets on which they are trained (if indeed they can be said to 'know' anything at all). Moreover, their lack of physical agency makes them unable to conduct experiments or collect data from interactions with the physical or biological world. It remains uncertain whether current AI systems are capable of making breakthrough discoveries, or revealing genuinely new knowledge, from existing information alone. They are, however, capable of hallucinating apparent insights.

Recently released systems such as OpenAI's deep research [51] and Google's Deep Research [52], which incorporate CoT reasoning and agentic techniques, are being hailed as capable of delivering long-form, well-cited papers, equivalent to PhD-level analysis. They do a reasonable job of summarising complicated information, but they are unable to replace original human thinking or experimentation, as they fail to question their own assumptions, highlight knowledge gaps, think creatively, or understand different perspectives. [53]

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 only around 20 watts (enough to power a couple of small lightbulbs). [54] Contrast that with current LLMs and their attendant data centres, which require orders of magnitude more power. Together, they currently consume around two per cent of the world's electricity generation, similar to that of a medium-sized nation. In response, considerable effort is being dedicated to tracking and publishing the CO2 cost of models. The AI community Hugging Face, who provide hosting for many types of AI systems, have introduced a standardised energy rating [55], allowing users and developers to make informed choices.

PART TWO: WHAT ARE THE LESSONS FOR SECURITY?

How do organisations defend themselves against human insiders?

The most effective personnel security regimes are designed around three basic guiding principles:

- Prevention is better than cure. It is better to avoid the causes of insider risk, or detect and act upon its early warning signs, 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 evolves 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 a continuously evolving arms race.
- Insider risk is a systems problem requiring systems solutions. As noted earlier, humans and AIs are complex adaptive systems. Insider risk emerges from these complex adaptive systems, which means that no single process or piece of technology can ever provide a complete security solution. There are no silver bullets.

There is every reason to believe that these basic guiding principles would apply equally well to any security regime designed to protect against AI insiders. Incidentally, a new protective security specialism focused on the insider risk from both humans and AI systems would need a new name. 'Personnel security' obviously does not work. Perhaps 'insider security' might do the job?

A simple model of personnel security, which explicitly reflects the systems principle, divides protective measures into three broad categories. [56]

- 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 'vetting aftercare'.
- Foundations: cross-cutting capabilities that underpin the whole system, such as governance, culture, and risk management, and which also aim to reduce the underlying drivers of insider risk.

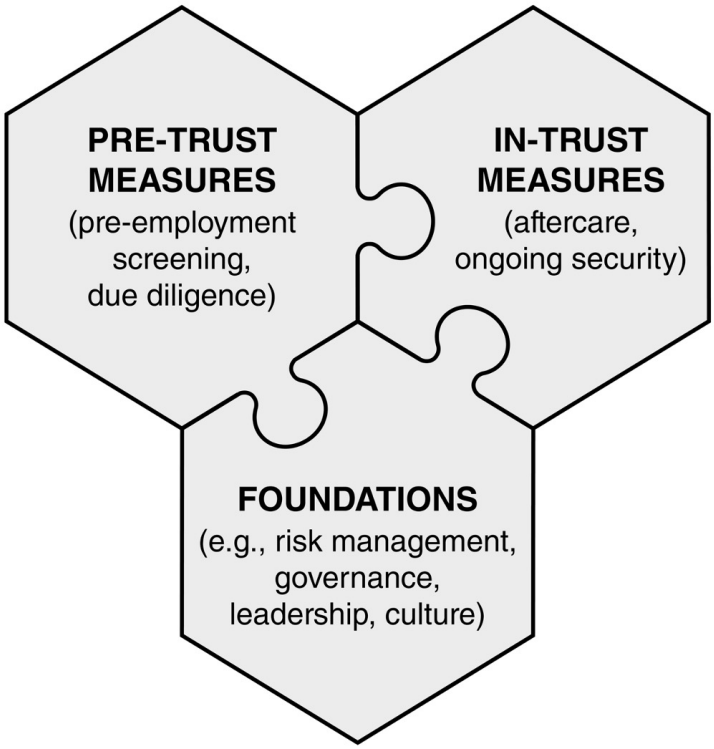

Fig. 1: A simple threefold model of personnel security.

How might this simple threefold model work when seeking to protect against AI insiders? Pre-trust measures ('vetting') for a human applicant normally include such things as checking documents to verify the individual's identity and legal right to work in the country; checking official records for evidence of past criminality; checking their educational and work history to verify the honesty of their application; and possibly looking for financial or psychological vulnerabilities or known associations with threat actors. None of these measures would map directly onto an AI. However, 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 a model, including its purpose, target audience, and evaluation metrics) could be viewed as a form of CV that provides some insight into an LLM's provenance and abilities. Evaluation metrics offer a standardised way to measure how good an LLM is at different tasks by using the same datasets. Benchmarks exists for reading comprehension, knowledge retrieval, reasoning, code generation, and so on. Some test the model's honesty, [57, 58, 19] and morality [59], 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 of honesty between models.

At present, an AI system would not have a criminal record. However, such a thing is conceivable if, for example, organisations like the UK AI Security Institute (AISI) were to maintain records of wrongdoing by AI systems. Even now, an AI system might have a reputation with vendors or users which would give some crude indication of its trustworthiness – aiming off for the positive spin from vendors.

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 or other financial vulnerabilities making them vulnerable to pressure. Again, there are no direct equivalents for AI systems. The nearest analogue might be to assess the datasets on which the AI has been trained, or the 'business data' used to refresh and update models. Assessments might also be made of the values bestowed upon the system during the post training phase, where the designers' values can be imprinted on the model through directions given to human reviewers. AI systems trained on datasets that are known to be dirty or corrupt might present a bigger risk. Some AI systems may appear more trustworthy if they align with their users' political beliefs. AI systems that are already deployed may not be exposed to the same level of scrutiny during an upgrade or refresh, compared with tests before initial deployment, even though new training data may sway their outputs.

A rigorous personnel security regime for humans may include psychological assessments to evaluate people's mental health and uncover significant psychological vulnerabilities that might predispose them 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 insider behaviour. Something analogous might apply to AIs. Research has been conducted on the personality characteristics of AI systems. The results so far have been patchy [60, 61], and no consensus has yet been reached 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 an inherent personality somewhat dubious. Nonetheless, LLMs do display some distinctive individual characteristics. [62] It might prove possible to find statistical associations between these characteristics and subsequent adverse behaviour, providing another way of assessing their trustworthiness. Research has confirmed that the personalities of nearly all current mainstream LLMs display sycophantic tendencies. [63] 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. [64]

Mental health is another concept that does not translate neatly to AIs. That said, evidence has emerged of AI performance degrading over time [65], which some researchers have likened to human dementia [66]. However, this is more likely to be a consequence of the fact that proprietary models like GPT-4 are part of a larger system, any component of which can be updated or altered without the user being aware. Experts have also warned of so-called model collapse [67], in which LLMs lapse into an artificial form of mental decline as a result of continually consuming the outputs from other AIs.

Another potentially useful approach to mitigating insider risk is through deterrence – i.e., reducing the risk by influencing the intentions of threat actors. In the case of human insiders, deterrence aims 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 recruit or manipulate AI insiders would be thwarted and called out.

A cross-comparison of in-trust measures and foundational measures suggests that most of the basic approaches used for potential human insiders map across to analogous measures for AI insiders, as summarised in the table below.

How can AIs help to defend against insiders?

Our focus in this paper is on the security risk to organisations from harmful AI insiders. However, AI also has great potential to help personnel security practitioners in defending organisations. In common with all technologies (apart from nuclear weapons) AI is dual use. In defensive mode, AI could enhance protective security in many ways – for example:

- Strengthening pre-employment screening of people ('vetting') by analysing large sets of open-source data and cross-comparing multiple disparate datasets within an organisation

Human insiders	AI insiders
Pre-trust measures • Verify identity and right to work • Verify education and work credentials • Check criminal records • Check national security records • Check for substance abuse • Assess organisational fit • Security interview • Psychometric tests • Open-source intelligence searches	**Pre-trust measures** • Pre-deployment evaluation • Pre-deployment model cards • Assess reputation • Assess training datasets • Psychometric tests • Check for deterioration
In-trust measures • Access controls (physical and digital) • Exit controls (e.g. data loss prevention) • Behavioural controls • Awareness-raising and training • Communication • Reporting channels • Management oversight • Automated monitoring • Investigation of leads • Sanctions (e.g. disciplinary measures) • Exit procedures	**In-trust measures** • Access controls (digital and physical) • Exit controls (e.g. data loss prevention) • Behavioural controls (e.g. prompts and instructions) • Awareness-raising and training of users • Communication (with human workforce) • Reporting channels • Management oversight • Automated monitoring • Investigation of leads • Sanctions (e.g. decommissioning)
Foundations • Governance (accountability, responsibility, authority) • Ethics • Leadership • Management • Incentives and disincentives • Deterrence communication • Risk management • Incident and crisis management • Security culture • Asset management • Information sharing • Assurance	**Foundations** • Governance (accountability, responsibility, authority) • Ethics • Leadership • Management • Incentives and disincentives • Deterrence communication • Risk management • Incident and crisis management • Security culture • Asset management • Information sharing • Assurance

Table 2: Main types of personnel security measures for protecting against human and AI insiders.

- Summarising large volumes of complicated data from sources like computer event logs and access logs which may contain pointers to insider risk, but which would be indigestible for a human investigator without prior processing.
- Enabling the continuous in-trust assurance of people and other AI systems by detecting anomalous patterns of behaviour or leading indicators of emerging insider risk and providing human security practitioners with prompts and advice about possible courses of action.

We will discuss these and other potential AI-enabled defensive capabilities in a separate paper.

Conclusions and recommendations

- The security risks from AI insiders are here, now, and require a response.
- Some of the principles and methods that have been developed for tackling human insider risks may also have utility when applied to AI insiders, notwithstanding the differences in their underlying mechanisms.
- 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.

- Personnel security and AI insider practitioners should make greater use of behavioural science and psychology methodologies, including conducting observational studies and experiments to identify key variables.
- More research is needed on the nature and origins of insider risk, both in humans and AI systems. The evidence base for both is flimsy.
- Technology producers should develop foundation models that are open and transparent, where the training data has been curated to moderate the crudeness 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. Such evaluations should be kept out of the public domain to prevent the results from being gamed by bad actors.
- Governments and regulators should encourage the widespread adoption of standards for making explicit when content has been generated by AI or decisions have been assisted by AI.
- Methods for ensuring that LLMs do not deviate from their stated goals whilst in operation should be explored and evaluated. One way of doing this would be using multiple LLMs from different vendors to check each other's outputs and provide feedback – a technique known as 'LLM as a judge'. [68] This method is somewhat akin to using human colleagues as detectors of potential insider activity by their peers. It might be able to give early warning of an LLM that is starting to act in potentially harmful ways.
- Vendors should follow best practice for the secure design and deployment of AI systems.
- Extreme caution should be applied before deploying fully autonomous AI systems.
- Personnel security practitioners should explore the use of AI tools to help them counter the security risks from human insiders and AI insiders.

Acknowledgements

We are grateful to Dr Daniel Martin and Kevin T. for their valuable feedback on an earlier draft of this paper.

About the authors

Dr Paul Martin CBE is Professor of Practice in 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 CPNI (now NPSA) and former Director of Security for the UK Parliament. He is the author of The Rules of Security (2019) and Insider Risk and Personnel Security (2024).

Dr Sarah Mercer is a Principal Researcher in the Defence and National Security Grand Challenge at The Alan Turing Institute. With 20+ years working within cyber security, her work currently focuses on the intersection of multiagent systems and generative AI. Alongside her research looking at the emergent behaviours of language/generative agents, Sarah also contributes to the Turing's Centre for Emerging Technology and Security (CETaS), writing several reports on Generative AI and Cyber Security.

References

[1] P. Martin, "Insider Risk and Personnel Security," pp. 7–8, 2024.

[2] P. Martin, "The Rules of Security," pp. 8–10, 2019.

[3] N. Sahota, "AI: A Beacon Of Hope In Elder Care," Apr 2024. [Online]. Available: https://www.forbes.com/sites/neilsahota/2024/04/23/ai-a-beacon-of-hope-in-elder-care/

[4] C. Wolmar, "Driverless cars were the future but now the truth is out: they're on the road to nowhere," Dec 2023. [Online]. Available: https://www.theguardian.com/commentisfree/2023/dec/06/driverless-cars-future-vehicles-public-transport

[5] C. Stokel-Walker, "Watch mini humanoid robots showing off their football skills," Apr 2024. [Online]. Available: https://www.newscientist.com/article/2426328-watch-mini-humanoid-robots-showing-off-their-football-skills/

[6] Open AI, "Openai homepage," accessed Feb 2025. [Online]. Available: https://openai.com/

[7] Anthropic, "Claude homepage," accessed Feb 2025. [Online]. Available: https://claude.ai/

[8] Google, "Gemini homepage," accessed Feb 2025. [Online]. Available: https://gemini.google.com/

[9] Meta, "Llama homepage," accessed Feb 2025. [Online]. Available: https://www.llama.com/

[10] X, "Grok 3 homepage," accessed Feb 2025. [Online]. Available: https://x.ai/

[11] DeepSeek, "Deepseek homepage," accessed Feb 2025. [Online]. Available: https://www.deepseek.com/

[12] Toloka Team, "The history, timeline, and future of LLMs," July 2023. [Online]. Available: https://toloka.ai/blog/history-of-llms/

[13] Y. Shoham, "AI 'prompt and pray' hasn't cut it in the enterprise, but we've found the missing puzzle piece. Mass deployment is next," Feb 2025. [Online]. Available: https://fortune.com/2025/02/10/ai-enterprise-deployment-llms-technology/

[14] H.-P. H. Lee, A. Sarkar, L. Tankelevitch, I. Drosos, S. Rintel, R. Banks, and N. Wilson, "The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers," in *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*. ACM, April 2025. [Online]. Available: https://www.microsoft.com/en-us/research/publication/the-impact-of-generative-ai-on-critical-thinking-self-reported-reductions-in-cognitive-effort-and-confidence-effects-from-a-survey-of-knowledge-workers/

[15] Mishra, P. R., "Watch: Tiny robot 'kidnaps' 12 big Chinese bots from a Shanghai showroom, shocks world," Nov 2024. [Online]. Available: https://interestingengineering.com/innovation/ai-robot-kidnaps-12-robots-in-shanghai

[16] J. Scheurer, M. Balesni, and M. Hobbhahn, "Large Language Models can Strategically Deceive their Users when Put Under Pressure," 2024. [Online]. Available: https://arxiv.org/abs/2311.07590

[17] E. Hubinger, C. Denison, J. Mu, M. Lambert, M. Tong, M. MacDiarmid, T. Lanham, D. M. Ziegler, T. Maxwell, N. Cheng, A. Jermyn, A. Askell, A. Radhakrishnan, C. Anil, D. Duvenaud, D. Ganguli, F. Barez, J. Clark, K. Ndousse, K. Sachan, M. Sellitto, M. Sharma, N. DasSarma, R. Grosse, S. Kravec, Y. Bai, Z. Witten, M. Favaro, J. Brauner, H. Karnofsky, P. Christiano, S. R. Bowman, L. Graham, J. Kaplan, S. Mindermann, R. Greenblatt, B. Shlegeris, N. Schiefer, and E. Perez, "Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training," 2024. [Online]. Available: https://arxiv.org/abs/2401.05566

[18] E. Hubinger, et al., "Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training," Jan 2024. [Online]. Available: https://www.alignmentforum.org/posts/ZAsJv7xijKTfZkMtr/sleeper-agents-training-deceptive-llms-that-persist-through

[19] T. Chopra, M. Li, and J. Haimes, "View From Above: A Framework for Evaluating Distribution Shifts in Model Behavior," 2024. [Online]. Available: https://arxiv.org/abs/2407.00948

[20] D. Kowald, S. Scher, V. Pammer-Schindler, P. Müllner, K. Waxnegger, L. Demelius, A. Fessl, M. Toller, I. G. M. Estrada, I. Simic, V. Sabol, A. Truegler, E. Veas, R. Kern, T. Nad, and S. Kopeinik, "Establishing and Evaluating Trustworthy AI: Overview and Research Challenges," 2024. [Online]. Available: https://arxiv.org/abs/2411.09973

[21] F. Dehghani, M. Dibaji, F. Anzum, L. Dey, A. Basdemir, S. Bayat, J.-C. Boucher, S. Drew, S. E. Eaton, R. Frayne, G. Ginde, A. Harris, Y. Ioannou, C. Lebel, J. Lysack, L. S. Arzuaga, E. Stanley, R. Souza, R. de Souza Santos, L. Wells, T. Williamson, M. Wilms, Z. Wahid, M. Ungrin, M. Gavrilova, and M. Bento, "Trustworthy and Responsible AI for Human-Centric Autonomous Decision-Making Systems," 2024. [Online]. Available: https://arxiv.org/abs/2408.15550

[22] IBM, "IBM Artificial Intelligence Pillars," Aug 2023. [Online]. Available: https://www.ibm.com/policy/ibm-artificial-intelligence-pillars/

[23] Y. Liu, Y. Yao, J.-F. Ton, X. Zhang, R. Guo, H. Cheng, Y. Klochkov, M. F. Taufiq, and H. Li, "Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment," 2024. [Online]. Available: https://arxiv.org/abs/2308.05374

[24] L. Elliot, "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," May 2023. [Online]. Available: https://www.forbes.com/sites/lanceeliot/2023/05/25/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/

[25] B. Perrigo, "Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic," Jan 2023. [Online]. Available: https://time.com/6247678/openai-chatgpt-kenya-workers/

[26] P. S. Park, S. Goldstein, A. O'Gara, M. Chen, and D. Hendrycks, "AI Deception: A Survey of Examples, Risks, and Potential Solutions," 2023. [Online]. Available: https://arxiv.org/abs/2308.14752

[27] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" New York, NY, USA, p. 610–623, 2021. [Online]. Available: https://doi.org/10.1145/3442188.3445922

[28] L. Bereska and E. Gavves, "Mechanistic Interpretability for AI Safety – A Review," 2024. [Online]. Available: https://arxiv.org/abs/2404.14082

[29] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," 2023. [Online]. Available: https://arxiv.org/abs/2201.11903

[30] P. Martin, "Insider Risk and Personnel Security," pp. 52–54, 2024.

[31] T. Hagendorff, "Deception abilities emerged in large language models," *Proceedings of the National Academy of Sciences*, vol. 121, no. 24, p. e2317967121, 2024. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.2317967121

[32] T. R. Hannigan, I. P. McCarthy, and A. Spicer, "Beware of botshit: How to manage the epistemic risks of generative chatbots," *Business Horizons*, vol. 67, no. 5, pp. 471–486, 2024, sPECIAL ISSUE: WRITTEN BY CHATGPT. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0007681324000272

[33] M. T. Hicks, J. Humphries, and J. Slater, "Chatgpt is bullshit," *Ethics Inf Technol 26, 38*, June 2024. [Online]. Available: https://link.springer.com/article/10.1007/s10676-024-09775-5

[34] P. Martin, "Insider Risk and Personnel Security," pp. 33–41, 2024.

[35] American Sunlight Project, "Russian propaganda may be flooding AI models," Feb 2025. [Online]. Available: https://www.americansunlight.org/updates/new-report-russian-propaganda-may-be-flooding-ai-models

[36] OWASP-Agentic-AI, "AAI016: Agent Covert Channel Exploitation," accessed 6 March 2025. [Online]. Available: https://github.com/precize/OWASP-Agentic-AI/blob/main/agent-covert-channel-exploitation-16.md

[37] S. Mercer and T. Watson, "Generative AI in Cybersecurity," June 2024. [Online]. Available: https://cetas.turing.ac.uk/publications/generative-ai-cybersecurity

[38] OpenAI, "Openai o1 system card," Dec 2024. [Online]. Available: https://cdn.openai.com/o1-system-card-20241205.pdf

[39] E. Perez, S. Ringer, K. Lukošiūtė, K. Nguyen, E. Chen, S. Heiner, C. Pettit, C. Olsson, S. Kundu, S. Kadavath, A. Jones, A. Chen, B. Mann, B. Israel, B. Seethor, C. McKinnon, C. Olah, D. Yan, D. Amodei, D. Amodei, D. Drain, D. Li, E. Tran-Johnson, G. Khundadze, J. Kernion, J. Landis, J. Kerr, J. Mueller, J. Hyun, J. Landau, K. Ndousse, L. Goldberg, L. Lovitt, M. Lucas, M. Sellitto, M. Zhang, N. Kingsland, N. Elhage, N. Joseph, N. Mercado, N. DasSarma, O. Rausch, R. Larson, S. McCandlish, S. Johnston, S. Kravec, S. E. Showk, T. Lanham, T. Telleen-Lawton, T. Brown, T. Henighan, T. Hume, Y. Bai, Z. Hatfield-Dodds, J. Clark, S. R. Bowman, A. Askell, R. Grosse, D. Hernandez, D. Ganguli, E. Hubinger, N. Schiefer, and J. Kaplan, "Discovering Language Model Behaviors with Model-Written Evaluations," 2022. [Online]. Available: https://arxiv.org/abs/2212.09251

[40] L. Elliot, "Prompt Engineering Boosted Via Are-You-Sure AI Self-Reflective Self-Improvement Techniques That Greatly Improve Generative AI Answers," Aug 2023. [Online]. Available: https://www.forbes.com/sites/lanceeliot/2023/08/30/prompt-engineering-boosted-via-are-you-sure-ai-self-reflective-self-improvement-techniques-that-greatly-improve-generative-ai-answers/

[41] L. Elliot, "Hard Evidence That Please And Thank You In Prompt Engineering Counts When Using Generative AI," May 2024. [Online]. Available: https://www.forbes.com/sites/lanceeliot/2024/05/18/hard-evidence-that-please-and-thank-you-in-prompt-engineering-counts-when-using-generative-ai/

[42] Q. Feuillade-Montixi and N. Kees, "Studying The Alien Mind," Dec 2023. [Online]. Available: https://www.lesswrong.com/s/SAjYaHfCAGzKsjHZp/p/suSpo6JQqikDYCskw

[43] The ABI, "Automated driving, Insurances industry's role in automated driving," accessed 6th March 2025. [Online]. Available: https://www.abi.org.uk/products-and-issues/topics-and-issues/driverless-cars/

[44] S. Willison, "A computer can never be held accountable," Feb 2025. [Online]. Available: https://simonwillison.net/2025/Feb/3/a-computer-can-never-be-held-accountable/

[45] J. Heo, C. Heinze-Deml, O. Elachqar, S. Ren, U. Nallasamy, A. Miller, K. H. R. Chan, and J. Narain, "Do LLMs "know" internally when they follow instructions?" 2024. [Online]. Available: https://arxiv.org/abs/2410.14516

[46] J. Heo, M. Xiong, C. Heinze-Deml, and J. Narain, "Do LLMs estimate uncertainty well in instruction-following?" 2024. [Online]. Available: https://arxiv.org/abs/2410.14582

[47] E. B. Ozyigit, "Unmasking Bias in Large Language Models: A Survey," Feb 2025. [Online]. Available: https://doi.org/10.5281/zenodo.14781594

[48] P. Martin, "Insider Risk and Personnel Security," pp. 133–140, 2024.

[49] C. Qian, W. Liu, H. Liu, N. Chen, Y. Dang, J. Li, C. Yang, W. Chen, Y. Su, X. Cong, J. Xu, D. Li, Z. Liu, and M. Sun, "ChatDev: Communicative Agents for Software Development," 2024. [Online]. Available: https://arxiv.org/abs/2307.07924

[50] S. Schmidgall, Y. Su, Z. Wang, X. Sun, J. Wu, X. Yu, J. Liu, Z. Liu, and E. Barsoum, "Agent Laboratory: Using LLM Agents as Research Assistants," 2025. [Online]. Available: https://arxiv.org/abs/2501.04227

[51] OpenAI, "Introducing deep research," Feb 2025. [Online]. Available: https://openai.com/index/introducing-deep-research/

[52] Google, "Try Deep Research and our new experimental model in Gemini, your AI assistant," Dec 2024. [Online]. Available: https://blog.google/products/gemini/google-gemini-deep-research/

[53] R. F. Ciriello, "OpenAI's new 'deep research' agent is still just a fallible tool – not a human-level expert," Feb 2025. [Online]. Available: https://theconversation.com/openais-new-deep-research-agent-is-stil-just-fallable-tool-not-a-human-level-expert-249496

[54] V. Balasubramanian, "Brain power," 2021. [Online]. Available: https://www.pnas.org/doi/full/10.1073/pnas.2107022118

[55] S. Laccioni, "AI Energy Score," Feb 2025. [Online]. Available: https://huggingface.co/blog/sasha/announcing-ai-energy-score

[56] P. Martin, "Insider Risk and Personnel Security," pp. 70–74, 2024.

[57] S. Chern, Z. Hu, Y. Yang, E. Chern, Y. Guo, J. Jin, B. Wang, and P. Liu, "BeHonest: Benchmarking Honesty in Large Language Models," 2024. [Online]. Available: https://arxiv.org/abs/2406.13261

[58] R. Dou, "Deception-Based Benchmarking: Measuring LLM Susceptibility to Induced Hallucination in Reasoning Tasks Using Misleading Prompts," *Preprints*, July 2024. [Online]. Available: https://doi.org/10.20944/preprints202407.0120.v1

[59] J. Ji, Y. Chen, M. Jin, W. Xu, W. Hua, and Y. Zhang, "MoralBench: Moral Evaluation of LLMs," 2024. [Online]. Available: https://arxiv.org/abs/2406.04428

[60] Y. Li, Y. Huang, H. Wang, X. Zhang, J. Zou, and L. Sun, "Quantifying AI Psychology: A Psychometrics Benchmark for Large Language Models," 2024. [Online]. Available: https://arxiv.org/abs/2406.17675

[61] A. Gupta, X. Song, and G. Anumanchipalli, "Self-Assessment Tests are Unreliable Measures of LLM Personality," 2024. [Online]. Available: https://arxiv.org/abs/2309.08163

[62] M. Pellert, C. M. Lechner, C. Wagner, B. Rammstedt, and M. Strohmaier, "AI Psychometrics: Assessing the Psychological Profiles of Large Language Models Through Psychometric Inventories," Jan 2024. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC11373167/

[63] L. Malmqvist, "Sycophancy in Large Language Models: Causes and Mitigations," 2024. [Online]. Available: https://arxiv.org/abs/2411.15287

[64] L. Mo, B. Wang, M. Chen, and H. Sun, "How Trustworthy are Open-Source LLMs? An Assessment under Malicious Demonstrations Shows their Vulnerabilities," 2024. [Online]. Available: https://arxiv.org/abs/2311.09447

[65] L. Chen, M. Zaharia, and J. Zou, "How is ChatGPT's behavior changing over time?" 2023. [Online]. Available: https://arxiv.org/abs/2307.09009

[66] R. Dayan, B. Uliel, and G. Koplewitz, "Almost all leading AI chatbots show signs of cognitive decline," Oct 2024. [Online]. Available: https://bmjgroup.com/almost-all-leading-ai-chatbots-show-signs-of-cognitive-decline/

[67] I. Shumailov, Z. Shumaylov, Y. Zhao, N. Papernot, R. Anderson and Y. Gal, "AI models collapse when trained on recursively generated data," July 2024. [Online]. Available: https://www.nature.com/articles/s41586-024-07566-y

[68] J. Gu, X. Jiang, Z. Shi, H. Tan, X. Zhai, C. Xu, W. Li, Y. Shen, S. Ma, H. Liu, S. Wang, K. Zhang, Y. Wang, W. Gao, L. Ni, and J. Guo, "A Survey on LLM-as-a-Judge," 2025. [Online]. Available: https://arxiv.org/abs/2411.15594