Lancaster University

Securing Agentic AI

WHITE PAPER

A collaboration between Lancaster University and Traversally

Traversally

Table of contents

> Trusting the agent to represent and aid an employee is one thing, but trusting the agent to represent you, as an employee, to your customers is another

Ian Makin,
Founder at Traversally

Securing Agentic AI

June 2026

Suppose you were looking for a holiday somewhere warm and sunny this July. If you met someone in the street for the first time, would you trust them with your bank details and let them book it for you? Probably not, and yet there are overlaps between this scenario and the use of agentic-AI systems to automate business processes. In this scenario, the agent 'receives' the customer payment information and context, and uses it to book their holiday. This is but one use case of many for agentic-AI, a technology that is being deployed globally by organisations attracted by its transformational capabilities. As the parallels in this holiday booking example highlight, however, trusting these systems does not come without risk.

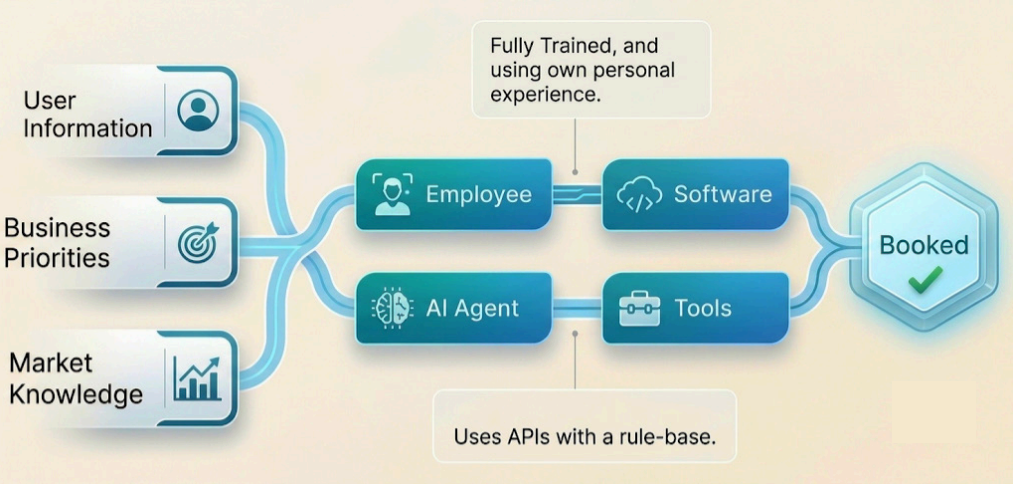


Figure 1 : This figure contrasts the paradigm of an automated AI agent booking a holiday using tools, versus a human you have never met booking a holiday using software. Naturally you expect a layer of trust between you and the human, and so why would you expect anything less from the AI agent?

Figure 1 shows an example high-level design for an agentic system. An AI Agent uses the contextual information available, and the set of tools it has access to, to achieve its goals. In normal operations, information about the agent's goals are set by the organisation deploying the system, and the agent trusts its aims and works towards them autonomously. But what if the instructions that governed their behaviour were hijacked by an adversary? Suddenly your agent is not acting how you intended: their goal, and now your organisation, has been hijacked. Before we look into this vulnerability in more depth, however, we first recall what Agentic-AI is.

What is Agentic-AI?

Generative AI systems, for example chatbots such as ChatGPT or Gemini, turn human prompts into desired output. Agentic-AI takes this further, consisting of AI agents that can plan, reason, and act autonomously to achieve some goals. To compare the two, generative AI is a support tool for a human to carry out tasks, whereas an AI agent can operate like a digital employee.

Securing Agentic AI

> A rogue agent acting against your goals is bad, but it gets worse when you realise that they are representing your organisation and brand and operating contrary to your values

Dr Phininder Balaghan,
Founder at Traversally

> Just because an organisation can use Agentic-AI to automate their operations, it does not mean that they should without carefully considering how this fundamentally changes their cybersecurity posture

Professor Nicholas Race,
Lancaster University

> Agentic-AI is an exciting technology, but it introduces new risks and an expanded attack surface that need managing carefully to ensure if can safely, securely, and effectively be deployed

Dr Edward Austin
Lancaster University

Why Agentic-AI?

In many cases, these agents will act independently, with limited or no human supervision, and they may work proactively to achieve their goals. This creates a highly attractive paradigm for organisations as agentic-AI can offer unparalleled efficiency for carrying out complex tasks, and they can operate at a velocity that creates significant potential for return on investment.

Real World Example

A simple and easily accessible AI application Traversally recently observed involved automated task list management. The agent was able to continuously reorder the tasks in which the company defined as the most productive opportunity first. This allowed the human operators to spend more time working on the highest potential, with the AI often demoting or even removing tasks as its research determined. They still worked through the list, but this showed AI optimizing their work and resulted in hitting targets much earlier each week.

But AI Agents Come With Risks

Despite the benefits for efficiency that agents offer, they also come with significant risks. For example, to achieve their goals and deliver impacts, agents necessarily need the freedom to act independently, however this creates the potential for them to make mistakes if there are not sufficient guardrails or human oversights in place. Imagine, for instance, an agent deciding that rather than read a database it should rewrite it.

Another risk, and one that is becoming well documented, is from prompt injection. This is where instructions are passed to the agent that cause it to behave in malicious, or at least unexpected, ways. For generative AI prompt injection can lead to harmful outputs, but for Agentic-AI it can lead to harmful actions. This problem is made worse by the fact that agents should be proactively ingesting data to inform their decisions: so now a new attack vector is created as a malicious actor does not need to touch the agent directly: they can cause harm by interfering with the external data that an agent uses to make its decisions.

Agents seem different, and they do expand your attack surface, but traditional cybersecurity principles need to apply to your system now more than ever as often these will lead to the exploits that will matter.

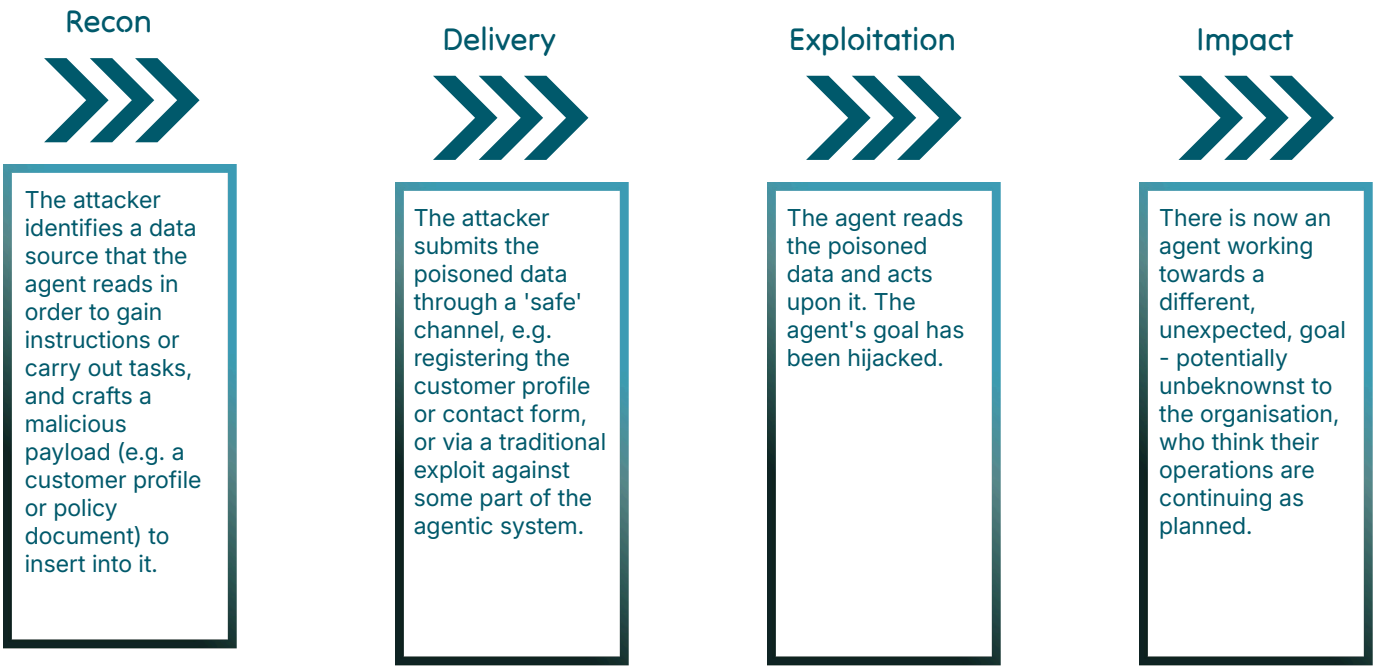
Securing Agentic AI

Technical Risk	What is it?	Strategic Impact
Indirect Prompt Injection	An adversary places malicious instructions in a document or data the agent handles.	Agent's guardrails are bypassed, allowing a threat actor to rewrite your operational procedures.
Insecure Output Handling	Agent generates actions or code that are carried out without being checked for malice.	An AI agent becomes an insider threat, able to harm or disrupt your system from inside your firewall.
Excessive Privilege When Calling Tools	The agent is given the ability to read and write files and APIs when it should be 'read-only'.	The agent is too independent. An error, or hijack, can result in permanent changes.
Data Ex-filtration	The agent is 'tricked' into sending sensitive data to an external recipient.	This will be damaging to the reputation of the organisation, and may lead to financial penalties.

Table 1 : Examples of technical security risks to an agentic system, and how they manifest into strategic impacts on an organisation.

Goal Hijacking

When deploying an agentic system the agents work towards a predefined set of goals. Often, these will be read in from other sources of information, no different to how teams across organisations learn of current strategies and priorities. In a goal hijack, however, these policies have been maliciously altered such that the agent is no longer working towards their original goal. The upshot of this is that the agent, faithfully following instructions, is not working outside of the expected strategy of an organisation - or worse, against it. This means that the highly efficient digital employee is now a rogue member of your team.



Securing Agentic AI

Bringing A Goal Hijack To Life

Imagine your organisation has just deployed a helpful new chatbot agent to manage the customer journey from sales to aftercare support. Like any employee, you give them access to the resources they need to complete their job. They can use internal systems, access customer and product details, carry out transactions, and receive training.

Unlike an employee, however, training for an agent takes the form of digital policy documents that instruct the agent how to behave. At first glance, this seems a benefit: with just a few sentences you can train your AI system to become a fully functioning digital employee. The problem, however, is that the documents and data that instruct your agent represent an attractive target to an attacker. By changing these instructions an adversary can influence the behaviour of your agent, hijacking its goals. For example, the customer agent may no longer act in your business's best interests, could offer customers refunds or unexpected sales, or could share private data with others. This means that your digital employee has become a rogue agent within your business.

The CAIRO (Cybersecurity Risks of Agentic-AI To Organisations) project, a partnership between Lancaster University and Traversally, is bringing to life the threat of goal hijacking through software demonstrators and thought leadership. These outputs show that, by only changing the content the policies an agent follows, harmful behaviour and malicious actions can be introduced without touching the AI model itself.

From Goal Hijack to Organisation Hijack

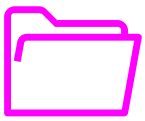
A rogue agent acting against your goals is bad, but it gets worse when you realise that they are representing your organisation and brand and operating contrary to your values. Now, instead of your agenda, they are working to someone else's. This means that the risk goes well beyond the financial and operational harms that a single hijacked agent can inflict, and moves into reputational damage. This is the worst case scenario; your agent has been hijacked, and so in turn has your organisation.

What Do I Do About This ?

Just because an agent can be fully autonomous, does not mean that they should be. In traditional cybersecurity users have their privileges restricted to control their access to systems, and for agents this goes further with both their accesses and their ability to act restricted. OWASP refer to this as "Constrained Autonomy".

In part, achieving this requires a secure-by-design approach to the agentic system, and ensuring that humans are strategically placed on the loop to provide oversight on the agent's actions. Equally crucial, however, is that these systems are carefully governed, treating the instructions for the agents with the same rigour as legal or financial documents, and ensuring that human co-pilots have clear decision-making protocols to follow. Whilst agentic-AI offers significant potential, it will not be the organisations with the fastest or most capable agents that receive the most value, but those that do not fall foul of the risks because they have the most securely governed systems.

Securing Agentic AI



Five Controls Every Agentic AI Programme Needs

- **Principle of Least Privilege**
  Restrict agents to only the data, systems, and actions required to perform their role.
- **Human-in-the-Loop Oversight**
  Introduce approval checkpoints for high-impact decisions and autonomous actions.
- **Policy Management Controls**
  Treat agent instructions and policies as controlled assets with formal governance and change management.
- **Continuous Monitoring**
  Monitor agent behaviour, tool usage, and outputs to quickly identify abnormal activity.
- **Segregation of Agent Duties**
  Avoid giving a single agent end-to-end authority over critical processes and transactions.

Securing Agentic AI

About the Authors

Dr Edward Austin is a research fellow at Lancaster University and has significant experience leading research on data-driven cybersecurity, and the safe and secure use of AI systems. He is the research lead for the CAIRO project, a collaboration between Lancaster University and Traversally.

Dr. Phininder Balaghan is the Co-Founder and CTO of Traversally, formerly the Global Head of AI for QinetiQ, a FTSE 250 organisation and a lead AI consultant for PA Consulting. He has led large-scale AI transformation across complex public-sector and enterprise environments. He has also advised UK government departments on AI strategy, governance, and assurance, and has represented the UK on international technology delegations. His work focuses on practical, risk-aware adoption of agentic and autonomous systems, helping leaders translate emerging AI capabilities into secure, governable, and operationally meaningful outcomes.

Ian Makin is a data strategist and two-time tech founder driven by using technology, specifically data, to solve mission-critical problems. Over a 25-year career, he has co-founded and scaled specialist data and AI governance consultancies, delivering secure-by-design solutions. Trusted by enterprise C-suites, Ian transforms organisational performance through secure, scalable, and trusted AI solutions.

Prof Nick Race is a Professor of Networked Systems at Lancaster University, the Associate Dean for Research for the Faculty of Science and Technology, and the Director of Lancaster's Academic Centre of Excellence for Cybersecurity Research. He brings extensive experience leading research projects on network security and automation, including the secure use of AI, and is the principal investigator for the CAIRO project.

Securing Agentic AI

About Traversally

Traversally provides governance, leadership, and assurance required to turn AI ambition into measurable business value - reducing risk, accelerating adoption, and ensuring AI investments deliver measurable ROI.

About Lancaster University

Lancaster University is a top 10 UK University, and world leading experts in AI and Cyber. They are recognised by as a centre of excellence, have significant funding through high award grants and industry partnerships.

About Cyber Focus

A huge thanks to CyberFocus for funding this work. The CyberFocus project has been designed to galvanise the North West cyber ecosystem by forging connections that instil confidence in research-led impact partnerships to propel national prosperity

This white paper has been supported by the PBIAA CyberFocus project, funded by the Engineering and Physical Sciences Research Council (EPSRC) [Grant No. EP/Z536052/1].