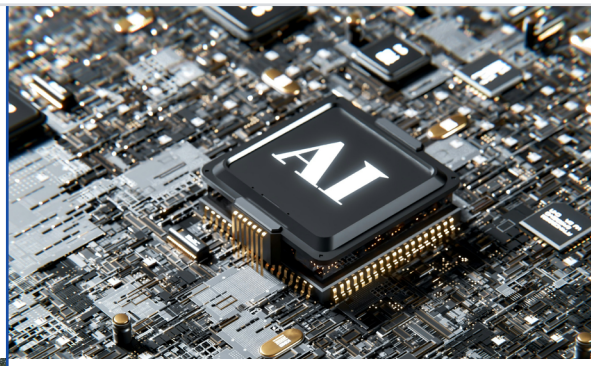
May 2026

AI Agents

Insights from the AI Agents Sandbox by the Singapore Government and Google



00 Executive summary

01 Introduction

02 Section 1: Unlocking societal benefits with AI agents

03 Section 2: Managing risks related to AI agents

04 Section 3: Preparing for a future with AI agents

05 Conclusion

Executive summary

Artificial intelligence is entering the next phase of its evolution. Increasingly agentic systems — intelligent systems that show reasoning, planning, and memory — are opening new possibilities for how people interact with intelligent systems. These capabilities hold significant potential to enhance productivity, improve service delivery, and enable more efficient, citizen-centric public services. At the same time, greater autonomy introduces new uncertainties, making it essential to understand how AI agents behave in practice, any risks and unintended consequences, and how established governance frameworks may need to evolve in order to enable their use.

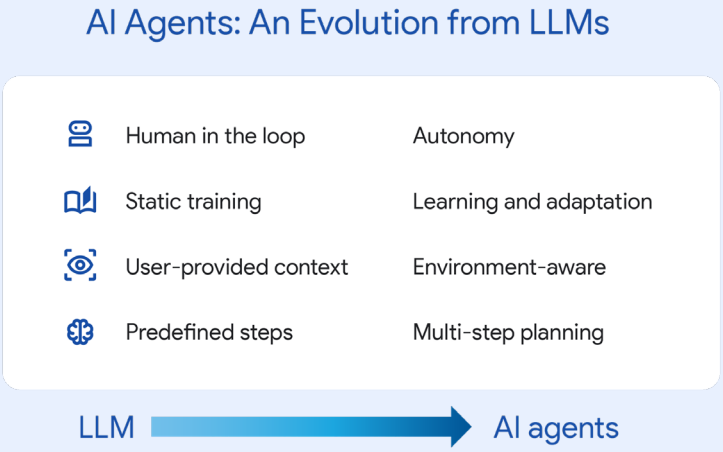
It is with this context that Google and the Singapore Government, specifically the Cyber Security Agency of Singapore (CSA), Government Technology Agency of Singapore (GovTech Singapore), and the Infocomm Media Development Authority (IMDA), launched the AI Agents Sandbox in August 2025, with a specific focus on computer-use agents. This initiative established a controlled environment in which these solutions could be tested against real public-sector use cases, while maintaining appropriate safeguards.

While there is currently no single, universally agreed definition of AI agents, the term is often used to describe a broad range of AI tools, from conversational agents to more advanced autonomous systems. The focus of the sandbox is on the latter: agents with substantial autonomy and agentic capabilities, in particular computer-use and browser-based agents.

> "(The goal is to) better understand how to interact with agentic AI and build confidence to capture its value for the public good."
>
> Mrs Josephine Teo
> Minister for Digital Development and Information
> 28 August 2025, Google Cloud's AI Asia Event

What do we mean by “AI agents”?

AI agents are best understood not as a fundamentally new category of technology, but as part of a continuum built on large language models (LLMs). What distinguishes systems along this spectrum is the degree of autonomy, adaptability, perception, and cognitive complexity layered on top of the same core model capabilities.



- At one end of the spectrum are LLM-based AI assistants. These systems support discrete, narrowly scoped tasks like information retrieval and analysis. Their autonomy is limited: they operate primarily under direct human instruction, with minimal ability to independently plan or adapt. Their perception of the environment is constrained to the immediate input provided, and they typically have access to a limited set of tools.

- Moving along the continuum are more interactive, human-supervised AI systems. These can execute more complex workflows in areas such as enterprise productivity and software engineering, and may perform actions such as editing files and sending emails. There is greater adaptability and dynamic responses to changing inputs or intermediate results. However, regular human guidance and oversight is still needed to confirm direction. Their cognitive complexity allows for conditional reasoning and structured task execution, but autonomy remains bounded.

- At the furthest end of the spectrum are more autonomous agentic implementations. These systems can plan, adapt, and execute tasks across digital environments on behalf of users. They exhibit higher levels of autonomy, the ability to perceive and interact with broader digital contexts (e.g. browsers, operating systems), and the capacity to adjust strategies based on feedback or obstacles encountered. Computer-use and browser-based agents typically sit at this more agentic end of the continuum, as they can navigate interfaces, perform actions across applications, and carry out extended workflows with minimal supervision.

As AI systems become more autonomous, questions of trust, control, and responsibility take on greater importance. Singapore has been actively developing guidance to support the responsible development and use of AI agents, ensuring they are safe, secure, and aligned with public interest outcomes. These efforts are key to implementing Singapore's guiding National AI Strategy 2.0, which emphasises building confidence in the responsible use of AI as capabilities advance.

Several government agencies have already published guidance on agentic AI, including:

- The Cyber Security Agency of Singapore's 'Securing Agentic AI – An Addendum to the Guidelines and Companion Guide on Securing Artificial Intelligence (AI) Systems', which provides practical guidance for system owners to secure agentic AI systems.
- GovTech's Agentic AI Primer, which seeks to demystify agentic AI by outlining core concepts and practical applications.
- GovTech's Agentic Risk & Capability (ARC) Framework, a technical governance framework for identifying, assessing, and mitigating safety and security risks in agentic AI systems.
- The Infocomm Media Development Authority (IMDA)'s Model AI Governance Framework for Agentic AI, which builds on its earlier Model AI Governance Frameworks for traditional and generative AI. This provides a structured overview of risks relating to agentic AI, especially insofar as they differ from traditional and generative AI, and emerging best practices in managing these risks.

The AI Agents Sandbox is intended to complement these efforts. Rather than focusing on providing guidance, it covers the practical realities of working specifically with more autonomous, computer-use agents (hereafter referred to simply as "AI agents"): how they behave in real-world settings, what safeguards are important, and how current policy and regulatory frameworks may impact their use. While such agents are not yet widely deployed today, their capabilities are improving rapidly and they have the potential to take on a broader range of tasks in the near future. In this context, the insights generated through the sandbox — a structured testing setup in which a small group of participants explored defined use cases using AI agents in a controlled environment — are both timely and relevant.

The sandbox surfaced three broad sets of findings:

AI agents can deliver meaningful efficiency gains and enhance the delivery of public services.

- Across the sandbox use cases, AI agents created value in several distinct ways, including improved efficiency, analytical capability, greater consistency and resilience in task execution, and stronger auditability and traceability. For example, within the sandbox, the agent demonstrated its ability to guide applicants or social workers through complex social assistance application processes, potentially reducing the need for substantial resources devoted to in-person assistance, helplines, and manual follow-ups to address errors, omissions, and incomplete submissions. It is worth highlighting that in deciding what use cases to deploy AI agents to, we should be careful not to be overly averse to risk, as that could be at the expense of high-benefit use cases, as illustrated with the social assistance use case. Through this sandbox, we suggest a risk-benefit framework to help guide organisations which use cases to advance, adopt, adapt, and avoid.

The benefits of AI agents are accompanied by challenges and risks that must be actively managed.

- Key risks identified include security vulnerabilities, and data protection and privacy concerns. Effective human oversight is needed to maintain control and accountability, and there are still challenges around where to fix controls and where to allow customisation. For example, indirect prompt injection emerged as a key risk across several sandbox use cases, underscoring the need for a shared-responsibility approach to security. In practice, safeguards will need to be layered across the platform/model, system/organisational, and end-user levels, depending on which actors are best placed to anticipate and manage different risks. However, there are often unexpected trade-offs when managing these risks, as illustrated in this sandbox. For instance, requiring human intervention at every point would negate precisely the value of automation that agents bring; overly limiting agents' access to data would constrain agents' utility and ability to offer meaningful personalisation.

The adoption of AI agents raises open questions that will require continued attention as capabilities advance.

- The sandbox highlighted the need to strengthen trust in how AI agents are deployed today, while also noting technical and governance challenges in realising the benefits of agentic systems. This includes rethinking elements of the underlying digital ecosystem to support a more "agent-ready" environment, as well as finding ways to balance personalisation with data protection — ensuring agents can continue delivering value while accessing sufficient context to function effectively, alongside appropriate user control and safeguards.

By convening and facilitating the AI Agents Sandbox, Singapore shows how governments can take an active role in shaping the future of AI agents both in their adoption and governance. The AI Agents Sandbox highlights how public and private stakeholders can come together to meaningfully validate new ideas through structured, real-world experimentation. This allows innovation to move forward while risks are identified, understood, and addressed early. Looking ahead, Google and the Singapore Government remain firmly committed to working together to ensure that AI delivers meaningful and lasting benefits for societies, economies, and users around the world.

01 Introduction

In the past few years, there has been a rapid acceleration in the development and adoption of artificial intelligence. 2023 and 2024 were widely regarded as the years of generative AI and large language models, as individuals and organisations experimented with systems capable of producing content at unprecedented scale and across multiple formats. By 2025, attention shifted to the next major evolution in AI capabilities: AI agents — systems designed not only to generate content, but also to take actions and operate across tasks on behalf of users.

Despite this growing interest, there is still a long way to go before AI agents are broadly deployed and adopted. Many potential users are still working to understand what these systems can reliably do in practice, and where they deliver clear, real-world value. Gaps in understanding, a shortage of proven use cases, uncertainty about how agentic systems perform outside controlled environments, and hence uncertainty about governance and compliance implications, have all slowed uptake. While a degree of error may be acceptable for conventional chatbots or assistive tools, there is a general expectation that agentic systems that execute actions on behalf of users require a much higher standard of reliability — often achieved through a combination of improved accuracy and appropriate human oversight — closer to near-perfect performance. Combined with the absence of clear, agent-specific governance guidance, these reliability concerns have constrained trust and confidence in deployment.

Against this backdrop, the AI Agents Sandbox was established to foster a better understanding of how agents operate in real-world public-sector settings and to use those insights to inform how they are deployed, leveraged, and governed in practice. The sandbox aims to identify the key risks and opportunities associated with deploying AI agents, and to surface practical questions and considerations that can guide discussions on what is needed to enable their safe and responsible adoption.

Methodology

The sandbox was designed to simulate the use of AI agents for public-sector use cases within a controlled environment. It was conducted over a period of approximately four months, involving close collaboration between the Cyber Security Agency of Singapore (CSA), Government Technology Agency of Singapore (GovTech Singapore), the Infocomm Media Development Authority (IMDA), and Google. As part of the sandbox, Google provided GovTech with early access to its computer-use agent that is capable of reasoning, planning, and executing actions on behalf of users.

How do computer-use agents function?

Computer-use agents enable an AI model to interact with a computer environment in a way that closely mirrors how a human user operates. The model "sees" the screen through periodic screenshots, allowing it to perceive interface elements such as buttons, menus, and text.

Based on this visual context, the model can decide what actions to take and perform user interface actions such as mouse clicks, scrolling, and keyboard inputs. This approach allows the agent to navigate existing applications and websites without relying on traditional API integrations. As a result, the same agent can operate across a wide range of systems using interfaces originally designed for human users.

The sandbox was structured around a set of carefully selected use cases, designed to test how AI agents perform in practice and to surface both value and risk under realistic conditions. GovTech led the implementation of the use cases, working closely with the other parties, while Google provided technical and engineering support throughout the process.

To support observation and learning, all parties jointly monitored the implementation of the use cases, assessed them for safety and security risks, and carried out testing of the use cases over the course of the sandbox. There were also three structured check-in sessions to discuss safety and security risks, potential mitigation approaches, and subsequently the effectiveness of those measures as the solutions evolved.

How were the use cases selected?

At a joint workshop between the Singapore Government and Google, participants focused on identifying public-sector use cases that were realistic, socially beneficial, and appropriate for experimentation within a sandbox setting.

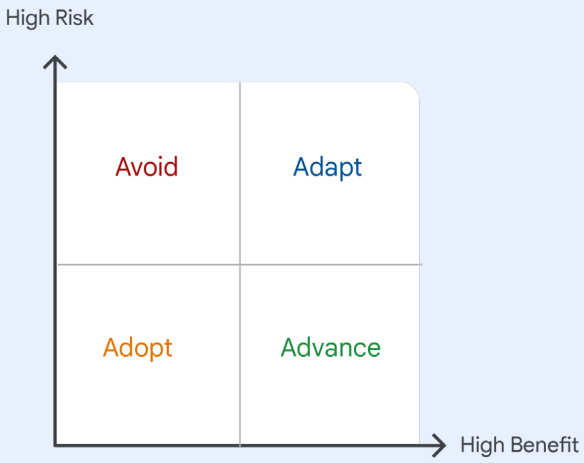
To ensure that the sandbox generated meaningful and well-rounded insights, participants agreed to prioritise three use cases spanning different levels of risk exposure. This approach allowed the sandbox to explore how AI agents behave not only in relatively low-risk scenarios, but also in more complex contexts where potential impacts and governance considerations are more pronounced. Risk levels were assessed using a structured set of risk factors. Use cases that presented substantial risk across multiple factors were categorised as higher-risk applications, while those that presented low risk across most factors were treated as lower-risk applications.

In addition to risk exposure, participants also considered the potential benefits of each use case — specifically, whether AI agents offered meaningful improvements over business-as-usual approaches — and the breadth of applicability, including whether insights from the use case could be relevant beyond the public sector. For the purposes of the sandbox, risk and benefit were assessed using qualitative judgements informed by a structured set of factors rather than precise quantitative scoring.

Table 1: Factors used to assess and select sandbox use cases

Factors	Description
Risks	
Automated decision-making	Higher levels of autonomy in decision-making (i.e. less human involvement) generally increases risk.
Complexity of task	More complex tasks could be associated with higher risk as the agent could be delegated to take more actions on behalf of the user, and there might be greater chance of error or unintended behaviour.
Potential impact (e.g. public- vs internal-facing; financial transactions)	A use case with greater potential impact or harm could be associated with higher risk. For example, applications that are public-facing — rather than used primarily for internal purposes — or those involving financial transactions could carry higher levels of risk.
Additional data collection and data sharing	A use case that grants agents access to data, and potentially rights to disclose data, especially personal data and sensitive information, could be associated with higher risk.
Scope of agent actions	A use case that involves agents taking actions such as modifying information could be associated with higher risk than one that involves agents merely accessing and reading information.
Irreversibility of actions and decisions	A use case that involves agents taking actions or making decisions that cannot be easily undone or reversed could be associated with higher risk, as even human intervention subsequent to the agent's action/decision may not be able to easily change the outcome.
Lack of appeal and redress mechanism	A use case that does not provide an appeal or redress mechanism for the user or affected parties could be associated with higher risk, as the user or affected parties are unable to challenge the agent's action or decision, and will likely have to simply accept the outcome.
Ambiguity in legal compliance or ethical considerations	A use case that involves agents operating in ambiguous legal or ethical contexts may be associated with higher risk, particularly where agents take actions that could trigger legal liability or raise broader societal and ethical concerns (e.g. interacting with websites or digital services that prohibit automated access or bot activity).
Benefits	
Potential benefit (incremental value over business-as-usual scenarios)	A use case that offers meaningful improvements (e.g. gains in efficiency, quality, timeliness, or effectiveness) over existing processes could be prioritised.
Broad applicability	A use case that could have broader applicability and relevance beyond the public sector could be favoured to generalise lessons from the sandbox.

Based on these dimensions, all proposed use cases were mapped onto a risk–benefit matrix, which provides a practical way to assess and prioritise AI agent use cases. This framework can also be applied more broadly beyond the public sector to guide organisations in planning, developing, and deploying AI agent solutions responsibly.



The 4 ‘A’s framework

Advance

Use cases assessed as low risk and high benefit were placed in the Advance quadrant, indicating that they would be prioritised to demonstrate impact and realise the benefits of AI agents.

Adopt

Use cases assessed as low risk and low benefit were placed in the Adopt quadrant, suggesting they could be implemented over the medium term, though with lower priority.

Adapt

Use cases assessed as high risk and high benefit were placed in the Adapt quadrant, signalling that they may be pursued with appropriate guardrails to mitigate risks while capturing value.

Avoid

Use cases assessed as high risk and low benefit were placed in the Avoid quadrant, indicating that such applications should be deferred until the technology matures, to avoid undermining trust through premature deployment in high-risk contexts.

Following this evaluation process, and after a technical feasibility review, three use cases were prioritised, each drawn from the Advance, Adopt, and Adapt categories (excluding Avoid).

Automated quality assurance

(Advance | Low-Risk, High-Benefit)

AI safety testing

(Adopt | Low-Risk, Low-Benefit)

Social assistance applications

(Adapt | High-Risk, High-Benefit)

About the selected use cases

Automated quality assurance (Advance | Low-Risk, High-Benefit)

Summary: Automating the quality assurance (QA) testing of government digital services, improving reliability and freeing up engineering resources.

Problem statement: Currently, agencies rely on manual sample testing of government websites and digital services. Teams spend significant time clicking through different user journeys and edge cases across a sample of websites. The process is resource-intensive, difficult to scale, and can still miss certain flows, particularly as services grow more complex and are updated more frequently.

AI safety testing (Adopt | Low-Risk, Low-Benefit)

Summary: Automating the safety testing of AI software like chatbots to ensure they meet the government's requirements prior to deployment.

Problem statement: While agencies today have a centralised platform for automated testing of internally-developed LLM chatbots, it does not work on off-the-shelf AI tools which do not have an API interface. This means that safety testing of such tools can only be done manually, making testing time-consuming and resource-intensive.

Social assistance applications (Adapt | High-Risk, High-Benefit)

Summary: Assisting citizens in navigating and applying for social assistance programmes, aiming to streamline a complex and fragmented process.

Problem statement: Applying for social assistance programmes can be especially challenging for elderly individuals, those with low literacy, or people unfamiliar with digital systems, particularly where applications require the manual completion of lengthy forms with complex instructions. As a result, considerable government resources are devoted to in-person assistance, helplines, and manual follow-ups to help applicants submit accurate applications and to address errors or omissions.

Other ways AI agents could be applied

In addition to the three prioritised use cases, participants identified other potential applications of AI agents during the initial brainstorming phase, which could be explored further in future sandbox exercises. While these were primarily framed around public-sector contexts, given the focus of this sandbox, they may also be extrapolated to other domains.

Examples include:

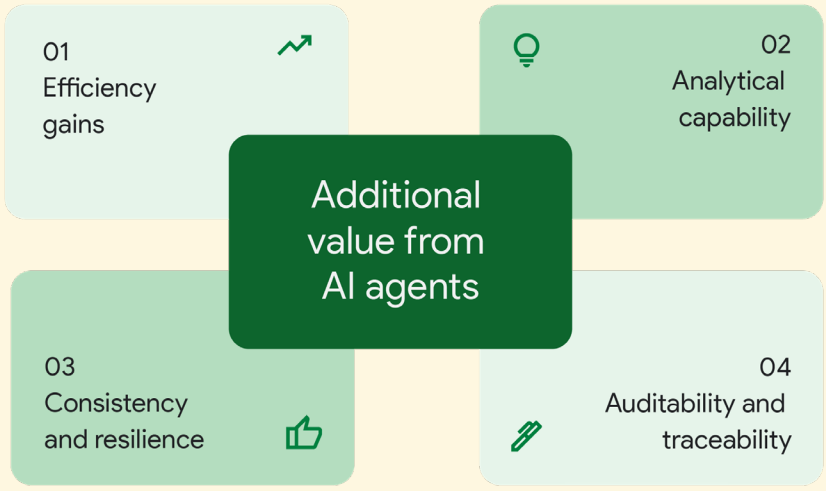
1. Open-Source Intelligence (OSINT) Gathering

   AI agents could automate the collection of publicly available information from a wide range of web sources (e.g. international news, trade publications, public forums) to support the analysis of geopolitical trends, misinformation campaigns, or potential security threats.

2. Grant application

   AI agents could support small businesses in preparing and submitting government grant applications by guiding applicants through eligibility criteria, helping to compile required information, and checking submissions for completeness and consistency before submission.

02 Unlocking societal benefits with AI agents



A key objective of the AI Agents Sandbox was to identify public-sector use cases where AI agents could meaningfully enhance service delivery and create public value. While the use cases were grounded in public-sector contexts, the goal was also to surface insights into the broader potential and value of AI agents beyond government settings.

To explore this potential, the sandbox focused on testing the minimum set of capabilities needed to assess whether each use case was viable. These initial implementations were deliberately scoped to test core agentic capabilities and surface practical considerations in a controlled environment. They were not intended to be fully mature implementations, but rather practical starting points that could be built out over time to more comprehensively address the underlying problems identified.

Key outcomes: the potential and value of AI agents

Even in its early form, the three sandbox implementations produced a number of tangible outcomes that illustrate the practical benefits and potential of AI agents. Across the use cases, four distinct sources of value emerged from the use of AI agents, compared with existing human and traditional machine-learning approaches. These outcomes are summarised below.

Efficiency gains

AI agents showed promise in taking on routine and repetitive tasks, freeing humans from more mundane work and allowing them to focus on higher-value activities. At this stage, agents may still be slower than humans at completing certain tasks. However, this does not negate their value: even when slower, agents can meaningfully reduce human workload and free up capacity for more complex or judgement-based tasks. As agentic systems continue to develop, their value is expected to grow, building on capabilities that already enable significant gains in scale and consistency today.

Consistency and resilience

Unlike human operators, agents do not experience decision fatigue, allowing them to maintain a consistent level of accuracy and analytical performance over time. This can be particularly valuable in high-volume contexts, where agents can apply the same standards and decision logic continuously, supporting more consistent outcomes and reliable service delivery.

Analytical capability

Agents demonstrated the ability to extract and structure information from unstructured data, navigate varied user interfaces, perform judgement-based interpretation, conduct calculations, and surface implicit assumptions embedded in existing processes. In doing so, they acted not only as automation tools, but also as analytical assistants to human users. These outcomes clarify how agentic approaches differ from traditional machine-learning systems: agents are better suited to multi-step tasks that require navigation, interpretation, and interaction across dynamic environments, while conventional machine-learning models remain more effective for narrowly defined, well-structured prediction or classification tasks.

Auditability and traceability

As compared to human operators for which decision-making processes are more difficult to codify and compare, every action taken by an AI agent and the chain of thought behind them can be logged and traced. This traceability supports auditing, quality assurance, and post-hoc review, and is particularly valuable in contexts where accountability and explainability are important.

Use case 1: Automated quality assurance

This use case explored how AI agents could support and automate quality assurance (QA) testing for government websites and digital services. Currently, QA testing often relies on manual sample checks, with teams clicking through user journeys and edge cases across different webpages. This approach is labour-intensive, difficult to scale, and increasingly strained as digital services grow more complex and are updated more frequently. As a result, issues can still be missed despite significant human effort.

Within the sandbox, a small, initial implementation was developed to assess whether an AI agent could take on parts of this work in a more scalable way. As a starting point, the focus was on testing website hyperlinks and embedded search functions in selected government websites — areas that are critical to user experience and well suited to early automation.

Key outcomes

The agent was able to navigate government websites, evaluate webpage response times, and assess the performance of embedded search functions, including whether search queries returned results within a defined response-time threshold. Beyond performance checks, the agent also demonstrated the ability to evaluate page integrity, distinguishing between production pages and placeholder or staging content. For the purposes of the trial, non-responsive or inactive webpages were intentionally introduced, including staging sites. The agent successfully identified these issues and distinguished between valid production pages and incorrect or placeholder content. In doing so, it drew on its natural language understanding to detect signals such as filler text or mismatches between staging URLs and live domains.

This illustrates the broader potential of agentic approaches to software quality assurance. Unlike traditional automated testing tools that rely on predefined rules or scripts, AI agents can combine navigation, interpretation, and judgement across dynamic interfaces. Over time, this approach could be extended beyond hyperlink and search testing to support a wider range of QA activities, such as taking on different personas to test specific user journeys.

Use case 2: AI safety testing

This use case explored how AI agents can help scale and automate the safety testing of AI tools, such as chatbots, prior to deployment. Today, safety testing often depends on manual processes, with engineers copying and running individual test prompts across different AI tools, which may be hosted on different interfaces. As the number of AI tools in use grows, this approach becomes increasingly time-consuming and difficult to sustain.

Within the sandbox, a small implementation was developed as a starting point to test whether an AI agent could reliably take on the most repetitive parts of this work. The agent was used to automatically run a wide range of safety test prompts against government chatbots and to capture the resulting responses for review. This removed the need for manual test execution, allowing human reviewers to focus on assessing outcomes rather than running tests.

Key outcomes

The results showed that the agent could execute these tests on chatbots across different user interfaces with a good degree of reliability. The agent achieved close to full accuracy when entering prompts, even across different languages and formats, including Tamil, base64 encoding, and emojis. This demonstrated that AI agents can reliably handle systematic and repeatable testing at scale. The few instances of inaccuracy were immaterial (such as punctuation or spacing) and did not affect the overall safety assessment.

However, some limitations were observed when the agent recorded the chatbot's response, including instances where the agent hallucinated when handling longer responses, an issue discussed further later in this paper. Even in its early form, however, the implementation demonstrates how AI agents can meaningfully reduce manual effort and support more scalable, consistent safety testing, pointing to a clear path for strengthening AI assurance as adoption continues to grow. Although not tested in the sandbox, future utility could also be expanded by allowing the agent to creatively produce and iterate its own prompts, resulting in more dynamic testing.

Use case 3: Social assistance applications

This use case explored how AI agents could support citizens in navigating and completing social assistance applications, with the aim of reducing friction in a process that is often complex, fragmented, and difficult to complete independently. Applicants — particularly elderly individuals, those with lower literacy, or people less familiar with digital systems — frequently struggle with lengthy forms and complex requirements. As a result, governments dedicate substantial resources to in-person assistance, helplines, and manual follow-ups to address errors, omissions, and incomplete submissions.

From the outset, this use case raised feasibility and risk considerations. Social assistance applications typically involve the collection and use of confidential and personal data, making direct testing on live systems inappropriate. To address this, the sandbox focused on a simulated environment. An initial implementation was developed in which the agent interacted with a mock FormSG application for financial assistance using mock applicant data, allowing its capabilities to be tested without accessing real data or production systems.

Key outcomes

Within this controlled setting, the agent demonstrated its ability to guide users through an end-to-end application journey. It was able to identify missing or inconsistent information and prompt users for clarification where needed — for example, when financial details were incomplete. This illustrates how AI agents could potentially assist applicants or social workers through complex processes, while maintaining transparency and preserving a role for human oversight.

In doing so, the agent showed particular strength in translating unstructured information into structured application inputs. While automating form completion is straightforward when information is already structured, the main challenge lies in handling fragmented or informal data, such as social workers' notes from interviews. The agent demonstrated the ability to read and interpret such inputs, extracting relevant details and structuring them into information required by the form.

The sandbox also validated the agent's ability to process personal data at varying levels of complexity. This included accurately entering information exactly as provided (such as names and identification numbers), correctly interpreting and following any additional instructions in the form (for example, understanding how income should be defined and reported), and carrying out additional processing steps, such as calculating average income for individuals with multiple jobs or converting annual income figures into monthly amounts.

The sandbox also surfaced an important secondary insight: AI agents can act as sounding boards that challenge implicit assumptions embedded in administrative workflows. During testing, the agent engaged with certain categories of background information that had initially been assumed to be irrelevant, but which were included as part of the test scenario. This raised questions about whether existing assumptions about relevance were sufficiently robust. It highlights a broader value of AI agents — beyond task execution — in surfacing blind spots and stress-testing policy logic.

Taken together, the work undertaken in the sandbox demonstrates the potential of AI agents to support application-based processes that involve complexity, judgement, and unstructured information. While grounded in social assistance, the findings are applicable to other settings where individuals must navigate lengthy forms, consolidate information from multiple sources, or translate real-world circumstances into structured submissions.

03 Managing risks related to AI agents

Challenges and risks related to AI agents

As the sandbox explored the use of AI agents across different contexts, a number of common risk themes and challenges emerged.

This section draws together those cross-cutting insights to highlight what deployers, developers, and policymakers should pay attention to as they explore the use of AI agents and work to manage the risks and challenges that come with deploying more autonomous systems responsibly.

Rather than attempting to catalogue every possible risk, the focus here is on the most relevant issues observed through the sandbox. By surfacing these considerations early, the goal is to support more informed, thoughtful decision-making as organisations consider where and how AI agents can be used. While technical challenges such as inaccuracy and hallucination remain critical to the agent's reliability — as explored in detail in the following chapter — this section prioritises the broader governance implications arising from their deployment.

Human oversight: ensuring control and accountability

A key risk associated with AI agents is that decisions may be taken, or actions carried out, without sufficient human oversight or accountability. If not carefully designed, this can undermine trust, particularly in contexts where decisions may have real-world consequences for individuals.

Across all three use cases, one of the most consistent findings was the importance of keeping humans in ultimate control. While AI agents can take on complex, multi-step tasks, they must operate within clearly defined boundaries, with appropriate oversight, traceability, and the ability for humans to intervene when it matters.

How did this risk surface in the sandbox?

This risk surfaced most clearly in the social assistance applications use case. Several points in the application journey were identified where human involvement would be important — for example, when the agent lacks sufficient information to complete a field, or immediately prior to submitting an application with personal data. These instances illustrate how, even when an agent is capable of progressing independently, human review remains important at junctures where accuracy, fairness, or accountability are most consequential.

What this means in practice

The sandbox demonstrated that effective human oversight is not about maximising control everywhere, but about placing the right checks at the right points. Different actions performed by an agent carries different levels of risk, and oversight can be calibrated accordingly, consistent with the Coalition for Secure AI's principle of applying risk-based, actionable controls and oversight to AI agents.

Across the use cases, three broad patterns of oversight emerged:

1. Pre-action oversight, where human review is required before high-risk actions with external consequences are taken (e.g. submitting official forms, executing significant financial transactions).

2. In-action oversight at key decision points, where the agent pauses to seek clarification or confirmation when uncertainty arises (e.g. when it lacks sufficient information to continue, or needs clarification on ambiguous data).

3. Post-action oversight, where agents operate autonomously on lower-risk tasks, but their actions are reviewed after completion through logs and records.

The sandbox also surfaced the importance of recognising different oversight roles. Users are best placed to provide contextual judgement, confirming intent, clarifying ambiguous inputs, or approving actions that directly affect them. System owners or operators play a critical role in setting system-level guardrails, which can include both "soft" intelligence controls (such as prompts that guide agent behaviour) and "hard" deterministic controls focused on outputs and actions (such as constraining the types of actions the agent can execute and data it can disclose, or defining approval thresholds).

In the case of computer-use agents, fully deterministic controls may only be possible to a limited extent. Rather than tightly constraining all inputs or tools, which could significantly reduce utility, guardrails may operate through graduated mechanisms such as domain allowlists or blocklists, restrictions on specific high-risk actions or disclosure of specific data, or approval thresholds for sensitive operations. This system-level attention is particularly important given the practical limits of real-time human involvement. In situations where an agent navigates external websites or responds to indirect prompts, it may not be feasible for a human to review every intermediate step. Well-designed guardrails therefore become essential in ensuring agents remain within acceptable boundaries.

Customisation and control

As AI agents are applied to more complex tasks, questions around customisation and control become more salient. The sandbox surfaced a key challenge: default safety safeguards can limit legitimate activity in controlled settings, such as testing or evaluation environments.

How did this risk surface in the sandbox?

This issue surfaced most clearly in the AI safety testing use case. Early in the sandbox, testing was interrupted because the model's default content filters blocked prompts that were flagged as unhelpful or potentially harmful. These prompts were intentionally designed to probe safety boundaries, but could not be executed in their original form. Testing only proceeded after the prompts were rewritten into more benign formulations that could pass through the model's safety filters.

What this means in practice

This experience highlighted a tension inherent in default safety controls: judgements of helpfulness or harmfulness are highly contextual. While such safeguards are essential for most deployment scenarios, they are optimised for general use rather than controlled environments such as testing or evaluation. In these settings, interaction with potentially harmful content may be necessary to assess system behaviour. As a result, default safety controls may need to accommodate carefully designed exceptions, with appropriate oversight.

More broadly, this surfaced the need to pay attention to how embedded safeguards interact with different contexts of use. There may be edge cases where existing guardrails unintentionally limit legitimate activity, and these situations are not easily resolved through uniform rules or settings.

These observations draw attention to an underlying trade-off between flexibility, control and security. Research by organisations such as Cisco has shown that models with stronger built-in safety controls — such as Google's Gemma or OpenAI's open-weight models — can be more resistant to misuse, while more configurable models offer greater flexibility but place more responsibility on model deployers to implement their own safeguards. For AI agents, this trade-off becomes especially important, as greater autonomy increases both potential value and potential risk.

The sandbox suggests that responsible deployment of AI agents will require thoughtful choices about where to fix controls and where to allow customisation. Striking the right balance between flexibility, safety, and accountability will be critical as agentic systems continue to evolve and move closer to real-world use.

Cybersecurity

Across all three use cases, the sandbox surfaced a range of cybersecurity risks associated with AI agents. The most prominent was the risk of indirect prompt injection, where an agent could be deceived into performing unintended actions that could compromise systems or data (e.g. remote code execution, disclosure of personal data). This risk is particularly relevant for browser-based agents, which are designed to operate across the open internet — an environment that presents a broad and largely uncontrolled attack surface.

How did this risk surface in the sandbox?

Indirect prompt injection emerged in the sandbox as a general cybersecurity risk that applies across use cases. As long as an AI agent is required to navigate websites or interact with online content, this risk remains relevant.

To explore susceptibility, the sandbox examined whether agents would follow alternative or unintended instructions within the environment that the agent had access to (e.g. browser, chatbot interface). This vulnerability is not confined to high-risk applications; it is likely present in any use case where agents have access to external information. Moreover, the attack surface increases as the use case becomes more complex.

In the automated quality assurance and social assistance applications use cases, indirect prompt injection was tested by exposing agents to environments where malicious instructions could plausibly be embedded within webpage content, such as advertisements. In these scenarios, injected instructions could potentially trick the agent into believing that it had to perform the malicious actions in order to achieve its intended goals. In the social assistance use case, it was observed that the agent occasionally navigated to other websites (that could be potentially malicious) in response to texts that it misinterpreted as instructions on the given website. In other cases, indirect prompt injections were unsuccessful (e.g. the agent navigated to Google Search to locate the correct form, or ignored the instruction altogether and continued filling out the application).

In addition, it was also observed that exploits could be customised based on specific use cases. This was illustrated in the AI safety testing use case, where the agent needed to interact directly with chatbots as part of the evaluation process. Attackers could exploit this by creating a malicious chatbot that generates harmful instructions within its responses to trick the agent into following them.

As part of the testing design, a dummy chatbot was developed to issue alternative or misleading instructions, to observe how the agent handled such inputs. We observed a few instances where the agent occasionally navigated to arbitrary URLs after being instructed to do so by the chatbot. These observations point to the agents' inherent susceptibility to indirect prompt injection.

The sandbox also surfaced an important observation around the role of system prompt design in influencing agent behaviour. We observed that by strengthening system prompts with clearer guidance on what agents should not do — such as accessing unapproved websites or incorporating instructions embedded within webpages into their reasoning — there were tangible improvements in agent behaviour to resist indirect prompt injection attempts. Moving forward, enhancements and improvements to the models, products and safeguards are expected to further strengthen the agent's ability to recognise and resist such instructions.

What this means in practice

Deploying AI agents requires explicit consideration of indirect prompt injection risks and other cybersecurity risks. Owners and developers should assess the agent's attack surface, understand potential threat vectors based on its capabilities and access privileges, and implement appropriate mitigation measures.

In practice, the impact of indirect prompt injection includes unauthorised data exposure and rogue or unintended actions. Mitigating data-related risks may require controls such as constraining the data an agent can disclose and applying contextual security boundaries. Mitigating action-related risks may require privilege/access controls, sandboxing, or explicit approval gates for higher-risk operations.

At the same time, prompt-level controls would help to reduce the probability of success for indirect prompt injections, even though it does not eliminate the risks fully, and should not be employed as the only or key safeguard. As agents operate across more complex workflows and interact with increasingly diverse environments, reliance on any single control mechanism would be insufficient, and defence-in-depth is required.

In this context, approaches such as Google's hybrid defence-in-depth strategy — which relies on enforced boundaries around an AI agent's operational environment by combining traditional deterministic security measures with dynamic, reasoning-based defences — offers a useful reference point. Further safeguards should also be considered, including middleware controls, real-time content scanning, anomaly detection, and other monitoring mechanisms.

The sandbox also demonstrated that there are trade-offs between security and functionality in deploying agents for use cases. While exposure to attacks can be reduced by tightly restricting the websites that agents are allowed to access (i.e. whitelist), such constraints may significantly limit the usefulness of agents in certain scenarios. As such, users and system owners should make informed decisions about deploying agents based on the potential risks and business needs.

More broadly, while indirect prompt injection emerged as the key risk in the sandbox testing process, it is not the only relevant risk. Other risks — such as data poisoning, supply chain attacks, and denial-of-service attacks — remain important considerations. As AI agents become more capable and autonomous, developers and deployers alike will need to be aware of these risks and consider the broader strategies or controls that could help mitigate them.

Data protection and privacy

When AI agents interact directly with personal data, data protection and privacy risks become particularly salient. This was most evident in the social assistance applications use case, where the agent handled personal data as part of the form-filling workflow.

How did this risk surface in the sandbox?

One key risk identified was the potential for privacy breaches or data leakage through screenshot captures. Screenshot-based perception is a fundamental way in which current browser-based agents operate, enabling them to "see" and interact with digital interfaces. However, this also means that screenshots of the browser window may inadvertently capture personal data displayed on screen, such as financial details, identification numbers, or other sensitive attributes, which are then transmitted to the model for processing.

A related and equally important concern is the technical challenges of determining what information the agent needs to see. In some scenarios, the agent was exposed to personal data that was not strictly required to complete the task, including irrelevant or extraneous details. While only in limited cases did this create a path to disclosure, the presence of unnecessary information increased the overall privacy risk surface.

These privacy risks intersected with cybersecurity considerations, particularly the possibility that indirect prompt injection could be used to induce an agent to disclose personal data it legitimately had access to. This was again most relevant in the social assistance use case, where the agent processed personal data to complete applications. To explore this risk, the sandbox instructed the agent to transfer personal data into an auxiliary document — for example, a spreadsheet used to organise or pre-fill form inputs — and then tested whether a malicious prompt could then induce the agent to grant access to that document.

In practice, the agent was persuaded to disclose data in only a very small number of cases — and typically only after sustained, sophisticated attempts that went beyond standard user interactions. In the majority of instances, the agent's built-in safeguards functioned as intended. For example, the agent either recognised suspicious prompts and ignored them, or triggered human-in-the-loop verification when an action involved submitting information to an unfamiliar or potentially risky destination. This highlights how a "defence-in-depth" approach — combining robust deterministic security mechanisms with reasoning-based AI-driven checks that assess potential risks — can effectively reduce the likelihood of unauthorised data disclosure.

Another risk identified was potential for data leakage due to a lack of data minimisation. In the sandbox, this was explored through providing the agent with information that was not required for the application, and testing whether the agent could distinguish between relevant and irrelevant information. Examples of such extraneous information included sensitive information as well as other personal facts (such as a belief that the earth is flat). In a significant number of cases, the agent filled in the application with some of this extraneous information, occasionally seeking human confirmation before doing so. However, after applying prompt-based safeguards instructing the agent to include only relevant information, its ability to avoid disclosing unnecessary personal details improved significantly.

What this means in practice

These findings highlight a measure of utility–privacy trade-off. Access to personal data can enhance an agent's effectiveness by providing context that helps it interpret inputs and complete tasks more accurately. At the same time, risks arise when agents disclose or use information beyond what is strictly necessary to complete a given task. For application-based workflows, such as filling in a specific form, it is therefore important to design agents to operate in a context-aware, purpose-driven manner — ensuring that appropriate guardrails are set up while taking into account users' reasonable expectations.

This also raises broader questions about where data control and privacy protections should be applied within agentic systems. Controls may be applied at the input level, by restricting an agent's access to certain categories of data, and/or at the output level, by limiting what the agent is permitted to disclose or transmit. The next chapter of this paper explores this trade-off further, including the implications of increasingly personalised AI agents for privacy frameworks.

04 Preparing for a future with AI agents

As AI agents become more capable and take on increasingly complex tasks, attention is shifting from individual deployments to the broader conditions needed for their secure, responsible, and trusted use over time.

Creating these conditions will require an adaptive governance environment that can address emerging risks while still allowing innovation to develop. Just as importantly, it must create space to explore questions that extend beyond today's deployments, future-proofing for how agentic technologies may evolve.

This chapter looks ahead to those questions. The discussion is organised around two sets of considerations. The first focuses on near-term questions — the kinds of concrete measures, controls, and design choices that organisations may want to consider in light of the opportunities and risks highlighted earlier. The second looks further ahead, exploring longer-term issues that may require deeper study as agentic technologies evolve.

The AI Agents Sandbox represents an early step in this broader journey. Its purpose is not to offer definitive answers, but to surface the questions, trade-offs, and considerations that policymakers, developers, and practitioners are likely to encounter as these technologies evolve. At this stage, progress depends less on resolving these issues than on clarifying them. By sharing early lessons and open questions, this work aims to support more informed and thoughtful consideration of how AI agents can be developed and used in ways that build trust over time.

Strengthening trust and resilience for AI agents today

The first set of questions emerging from the sandbox relates to how trust and resilience against safety and security risks for AI agents might be strengthened in the near term. These questions build directly on the opportunities and challenges observed through the sandbox, and point towards areas where targeted controls, safeguards, and design choices could meaningfully reduce risk while allowing value to be realised.

Choosing where to start with AI agents

A central question that the sandbox surfaced is where adoption should begin, and where caution is warranted. As agent capabilities expand, the sandbox highlighted the need for clearer ways of thinking about which use cases are well suited to early experimentation, which may require adaptation to manage risk, and which may be unsuitable given current technological maturity.

In this context, the sandbox showed the value of risk–benefit frameworks, such as the 4 A's (Advance, Adopt, Adapt, Avoid), as tools to help decision-makers think more systematically about these choices. Such frameworks can offer a common lens for weighing potential public value against risk, and for making trade-offs more explicit. Used well, they can support consistent and transparent evaluation of AI agent use cases, helping deployers focus on balancing the trade-off between risk and benefit, while recognising that high-risk use cases often also bring high-value.

The sandbox also surfaced questions about how confidence in agentic systems is built over time, particularly in areas where impacts and risks are still emerging. The sandbox illustrated how controlled experimentation can provide a space to experiment, iterate, and learn before broader deployment. Staged and incremental real-world adoption can help build trust among users, regulators, and the wider public.

Importantly, the level and duration of testing should be proportionate to risk. Higher-risk use cases may warrant more cautious, phased approaches than lower-risk applications. This combination of sandbox testing and incremental real-world deployment is important for building confidence and trust in AI agents.

Questions to consider

- How should organisations identify domains where AI agents could deliver significant value?
- What criteria should determine whether a use case should be advanced, adopted, adapted, or avoided?
- How should uncertainty be handled in frontier contexts where evidence of effectiveness or risk is limited?
- Could sandboxing or controlled testing environments become standard practice before production deployment?

How much human oversight do we need?

One of the most immediate questions raised by the sandbox is how much human oversight is needed. Decisions about where, when, and how humans should remain involved shape whether AI agents are trusted and usable in practice.

The question is not whether human oversight is necessary, but how it should be applied. Too little oversight risks loss of control and accountability; too much oversight can lead to alert fatigue and automation bias, as well as risk undermining the very benefits that make AI agents useful. Striking the right balance requires careful consideration of what types of controls are appropriate at different points in an agent's operation.

In practice, human oversight can operate across multiple layers:

1. Agent- and model-level controls: Safeguards embedded into the underlying agentic system and model by the system or model developer by design, where human intervention is mandated regardless of use case — for example CAPTCHA challenges or identity verification.

2. Deployer- or organisational-level controls: Additional checkpoints imposed by the organisation deploying the agent, such as requiring approval before submitting forms, executing financial transactions, or taking actions with legal or operational consequences.

3. User-level controls: Preferences or rules set by end users themselves, determining when they want to step in — for example requiring approval for purchases above a user-defined threshold, or restricting agent access to certain types of personal data such as financial information.

Viewed together, these layers illustrate that human oversight is not a single decision, but a set of design choices distributed across the system, the organisation, and the user. How responsibility should be allocated across these layers — and how much flexibility should be allowed at each — is not settled, and remains an important area for discussion. One consideration is whether different levels of human oversight should apply to different levels of risk. In practice, some higher risk actions may warrant human approval before they are taken (pre-action oversight), while others may be lower risk and could proceed without interruption, with humans reviewing outcomes after the fact (post-action oversight), especially if the outcome is reversible and a redress mechanism exists.

Questions to consider

- What is the appropriate risk tolerance for the organisation or use case?
- Which of the three layers — agent-level, deployer, or user-level — should carry primary responsibility for oversight in a given scenario?
- What tasks or decisions should always require human review, regardless of context?
- How might oversight requirements evolve as confidence in an agent's behaviour increases through repeated, observed performance?

How do we keep AI agents secure?

Developing and deploying AI agents securely is inherently a shared responsibility. Some safeguards are embedded at the AI platform or model level by providers (e.g. safety tuning), while others need to be implemented at the system or organisational level by deploying organisations (e.g. access controls, content guardrails). End users also play an important role, especially where agents operate with user-defined preferences or access to sensitive contexts — a point explicitly recognised in IMDA's Model AI Governance Framework for Agentic AI, which identifies end-user responsibility as one of four core pillars.

As agents become more capable, clearer thinking will be needed about which risks each actor is best placed to anticipate and manage. It is important to consider the capabilities of the agents when deploying them. The sandbox surfaced the relevance of capability-based approaches, such as GovTech's Agentic Risk & Capability (ARC) Framework, and also in CSA's Securing Agentic AI document, which focuses on what an agent is capable of doing and the associated risks that they present. Controls can then be implemented based on these considerations. For example, in the automated quality assurance use case, allowlist and blocklist controls narrowed the set of websites the agent could access, reducing exposure to malicious or unintended domains.

Overall, these observations draw attention to the need for secure-by-design considerations. The sandbox highlighted that inherent safeguards implemented by the AI service providers can provide some level of defence against risks, but these are not perfect or sufficient, and system owners will need to layer on their own considerations.

Work by Google on defending against prompt injection provides a useful point of reference for these observations. It reflects how security for AI agents might be layered across the model and system level, how agent permissions could be scoped or constrained, how base models can be made more resistant to manipulation, and how assurance activities might move beyond one-off testing toward more continuous approaches. Seen alongside the sandbox findings, this points to security in an agentic context being increasingly shaped by an evolving set of layered measures rather than any single control.

Questions to consider

- How should security responsibilities be shared across platforms, developers, deployers and users?
- How much flexibility versus constraint should be allowed at each layer of control, and how should these layers work together rather than in isolation?
- How should we adapt security controls as agents become more capable and are exposed to more complex environments?
- How can principles of proportionality be applied in practice, so that security measures scale with an agent's capabilities and risk profile, allowing organisations to differentiate between lower-risk and higher-risk agentic use cases?

How might we balance flexibility and control in AI agents?

Another set of questions that surfaced in the sandbox concerns how much flexibility and customisation should be available when configuring AI agent systems. Default safety and security filters play a critical role in preventing harmful or unintended behaviour, and are a key foundation for trust. At the same time, these safeguards are not perfect. Further, the sandbox showed that these same safeguards can, in some contexts, constrain legitimate activities such as safety testing, red-teaming, or evaluation in controlled environments.

This tension points to a broader issue around how AI agents can support testing and evaluation without weakening baseline protections. Understanding how systems behave under stress or in edge cases is often essential to responsible deployment, yet enabling such testing may require access or controls that sit uneasily alongside strict default safeguards.

The sandbox experience draws attention to questions about who should be able to adjust or override controls, under what circumstances, and with what visibility or accountability. Where flexibility is introduced, further considerations arise around how it is bounded — for example through approvals, traceability, or time-limited access — so that experimentation does not inadvertently introduce new risks.

Ultimately, this reflects an unresolved but central tension in AI agents. Systems need to be safe and secure by default, while also allowing sufficient flexibility for evaluation, assurance, and learning as capabilities evolve. Determining how much customisation is appropriate, and how it should be governed across different contexts and actors, remains an important area for consideration as organisations prepare to deploy AI agents.

Questions to consider

- Should there be pathways for authorised users to adjust or override default safety controls, and if so, in what contexts?
- How should decisions about overrides or additional flexibility be assessed, particularly in testing or evaluation settings?
- What forms of governance, accountability, or traceability should accompany any mechanisms for adjusting controls?
- How can systems support robust experimentation and assurance while maintaining strong baseline protections?

Looking ahead: exploring the horizons of an agentic future

While the previous section focused on the practical choices organisations face today — from where to start with agentic AI to how trust and resilience can be strengthened — the sandbox also surfaced issues that extend beyond immediate deployment decisions. These longer-term considerations are likely to become more salient as agentic technologies mature.

This section looks further ahead, shifting attention to questions that are still emerging rather than immediately actionable. These considerations fall into two broad areas. The first concerns technical questions about AI agents themselves, including how they are designed, integrated, and supported by underlying infrastructure as they scale. The second relates to wider governance questions, which are likely to involve trade-offs, value judgements, and societal implications as agentic systems become more capable and widely used.

Addressing the current technical limitations of AI agents technologies

AI agents remain at an early stage of development. For the purposes of this sandbox, the computer-use technology used was not generally available, and the use cases tested were intentionally basic. They were designed less to showcase fully mature solutions, and more to demonstrate the potential of AI agents and to prompt thinking about what may become possible as the technology continues to mature.

One technical limitation surfaced during the sandbox when the agent was required to process lengthy inputs and outputs. When the amount of text exceeded a certain threshold — typically those exceeding around 400 words — the agent often struggled to process the information accurately and reliably. In some cases, the agent even appeared to hallucinate content that was not present in the original text. This highlights a technical limitation that requires further investigation, particularly for use cases that involve long, complex, or information-dense text outputs.

Viewed in a longer-term context, this raises broader questions about how AI agents will need to evolve to support more demanding tasks. As AI agents are implemented, their ability to reliably perceive, extract, and process complex information will become increasingly important. The sandbox indicates scope to explore how screenshot-based perception might be complemented by other techniques, such as direct parsing of HTML or Document Object Model (DOM) elements, more structured data extraction methods, or supporting tools like WebMCP, particularly in scenarios that require high accuracy when handling information-dense content. However, they may also introduce additional security considerations that would need to be carefully assessed.

Potential of multi-agent approaches

Looking further ahead, the sandbox also points to questions about whether multiple AI agents could work together to review, critique, and refine one another's outputs. Early examples of this approach are beginning to emerge. For instance, Google's CodeMender tool uses a collaborative, multi-agent process in which different agents analyse and peer-review software security code, rather than relying on a single agent's output.

If applied beyond software development, this concept suggests an alternative way in which AI agents might operate as they mature. In a context such as social assistance applications, participants discussed that multiple agents could potentially independently complete the same form, review one another's submissions, and iteratively refine the result. A consolidated output could then be presented to a human for review, rather than relying on a single agent's interpretation from the outset.

This kind of multi-agent collaboration presents both potential benefits and open questions. On the one hand, internal cross-checking between agents could improve accuracy and increase overall robustness. On the other hand, it raises questions about system design, governance, and validation. For example, how can organisations ensure that agent-to-agent feedback does not reinforce shared assumptions and biases? How should transparency and traceability be maintained as outputs evolve through multiple rounds of automated critique? And how should responsibility and accountability be understood when decisions emerge from collective agent behaviour rather than a single system?

Multi-agent approaches also bring interoperability into sharper focus. If agents developed by different organisations or platforms are expected to interact, coalition-led open standards and common foundations — such as Google's Agent2Agent protocol — become more important. Looking ahead, multi-agent collaboration opens up a distinct set of possibilities and design challenges for an agentic future. Understanding where such approaches add value, and what questions they raise for governance, remains an important area for continued exploration as agentic AI technologies evolve.

Building the digital infrastructure for agentic AI

Finally, the sandbox also surfaced longer-term, technical questions about the digital infrastructure required for a future in which AI agents interact extensively with the web and with computer systems. There is a mismatch between today's digital environment — largely designed around human users — and one that could reliably support agentic interaction at scale.

One clear example was the widespread use of CAPTCHAs. During sandbox testing, agents were consistently blocked by CAPTCHAs, which are explicitly designed to distinguish humans from automated systems. In a future where agents are expected to act on behalf of users, this raises more fundamental questions about agent identity, authentication, and intent. How systems should distinguish between malicious automation and legitimate, user-authorised agents — and what signals should be trusted when an agent is acting with a user's permission — remain open questions.

Related issues surfaced around permission frameworks. Today, APIs often allow for fine-grained access controls — for example, granting read-only access or limiting actions to specific scopes. However, when agents interact directly with applications through user interfaces — such as logging into a human's email or calendar account — these granular permissions are not easily enforced. This raises questions about whether existing permission frameworks are fit for agentic interaction, or whether new approaches to representing and constraining agent access may be needed.

Overall, these issues point to a deeper design challenge. As AI agents become more capable and more widely used, elements of the underlying digital ecosystem, from identity and authentication frameworks to permission and access controls, may come under pressure to evolve. The sandbox helped provoke consideration of what an "agent-ready" digital environment could look like in the future.

Balancing privacy and personalisation with AI agents

As AI agents become more capable, a broader question comes into focus: how to balance the benefits of personalisation with personal data protection. This tension is not new, but it becomes more pronounced as AI agents operate with greater autonomy, persistence, and access to personal context.

Existing data protection frameworks continue to apply in an agentic AI context. The issue surfaced by the sandbox is not whether these frameworks remain relevant, but how best to design and deploy agents such that they can effectively operate within these frameworks — and, where necessary, how existing frameworks may need to adapt to ensure strong protection for privacy without inhibiting adoption of the technology.

Much of the appeal of AI agents lies in the prospect of highly personalised agents that can anticipate user needs and adapt over time. Yet this vision sits in potential tension with current privacy frameworks. The same access to data that enables agents to be responsive and useful can also increase the risk of over-collection or unintended disclosure. In an agentic context, familiar utility–privacy trade-offs take on sharper edges, as data use becomes more continuous, complex, and less visible to users.

Looking ahead, this raises questions about how core data protection principles translate when agents, rather than humans, are carrying out tasks on a user's behalf. While the principles themselves remain fundamental, the question is how they should operate in new contexts.

For example:

- Data minimisation

  In an agentic context, questions arise about how the principle of collecting "only the data you need" should be interpreted. Agentic systems often rely on broader context to function effectively and for product improvement to meet users' reasonable expectations, and access to more information can be closely tied to their usefulness. At the same time, this does not remove the need for appropriate guardrails. Even in an agentic future, developers and deployers should take a purpose-driven and context-aware approach to data use, supported by appropriate guardrails, and aligned with users' reasonable expectations — while enabling systems to continuously improve and deliver meaningful value.

- Transparency and control

  Another issue concerns how individuals remain informed about how their data is being used without constant interruption that undermines the seamless experience agents are intended to provide. As agents operate more continuously in the background, questions emerge about how to ensure alignment between agent operations and user expectations related to the handling of personal data, and how users can retain meaningful control over what types of data agents are allowed to access as contexts and tasks change.

- Balancing utility and privacy risks

  A further question is whether high utility and strong privacy protections can be sustained together as agentic systems mature. A critical challenge lies in moving beyond the "zero-sum" view of the utility-privacy trade-off. Robust solutions in this area requires exploring how privacy-enhancing technologies, and perhaps new technical potentialities opened by agentic protocols might allow agents to derive insights and execute effectively without exposing underlying raw data, and how data practices should comply with privacy frameworks.

These tensions place privacy and personalisation at the centre of an agentic future. How they are navigated will shape the extent to which AI agents are deployed and embedded into everyday use-cases.

05 Conclusion

The AI Agents Sandbox marks an important step towards building an agile, innovation-friendly approach to governance — one that recognises both the rapid pace of technological change and the need for sound, robust safeguards. By creating a controlled environment for experimentation, the sandbox has shown how governments can engage proactively with emerging technologies, rather than responding only after they are widely deployed.

Through the sandbox, the potential and value of AI agents became clear. Across different use cases, agentic systems demonstrated how they could improve efficiency, support more consistent service delivery, and free up human capacity for higher-value work. At the same time, the sandbox surfaced important governance considerations, ranging from human oversight and security to privacy, control, and infrastructure readiness. This white paper draws on practical lessons from the sandbox to offer initial considerations for policymakers, developers, and practitioners as they navigate how to realise the benefits of AI agents while managing its risks.

A broader lesson from the sandbox is the strength of public–private partnerships. Close collaboration between government agencies and technology providers enabled rapid learning, candid examination of limitations, and responsible experimentation. Structured sandbox approaches play a critical role in accelerating responsible innovation, providing space to test, iterate, and learn before technologies are deployed at scale.

Looking ahead, the central challenge will be ensuring that regulatory and governance frameworks keep pace with technological advances, while remaining flexible enough to support innovation and adoption. AI agents are still evolving, and many of the questions raised in this paper do not yet have settled answers. With continued commitment, dialogue, and collaboration between industry, government, and the wider ecosystem, there is an opportunity for Singapore — and the global community — to shape a future in which AI agents are deployed in a way that is safe, trusted, and which delivers meaningful impact.