自主智能体人工智能模型治理框架

新加坡资讯通信媒体发展局 (Infocomm Media Development Authority, IMDA)

2026年1月

中文翻译版 · 仅供参考，以英文原文为准

原文来源：https://www.imda.gov.sg/-/media/imda/files/about/emerging-tech-and-research/artificial-intelligence/mgf-for-agentic-ai.pdf

---

目录

执行摘要

自主智能体AI（Agentic AI）是人工智能的下一次进化，对用户和企业具有变革性的潜力。与生成式AI（Generative AI）相比，AI智能体（AI Agent）能够采取行动、适应新信息，并与其他智能体和系统互动，代替人类完成任务。尽管应用场景仍在快速发展，但智能体已经在通过编码助手、客服智能体和企业生产力工作流自动化等方式改变职场。

然而，更强大的能力也带来了新的风险。智能体对敏感数据的访问权限以及对其环境做出变更的能力——例如更新客户数据库或执行支付——都是一把双刃剑。随着我们逐步部署具有复杂交互能力的多个智能体，其结果也变得更加不可预测。

人类必须保持问责并妥善管理这些风险。虽然现有的可信AI治理原则——如透明性、问责制和公平性——仍然适用，但需要在实践中针对智能体进行转化。有意义的人类控制和监督需要被整合到自主智能体AI的全生命周期中。然而，必须在持续的人类监督与规模化运营之间取得平衡，因为对所有智能体工作流的持续人工监督在规模化时变得不切实际。

自主智能体AI模型治理框架（MGF） 为组织提供了关于自主智能体AI风险及新兴最佳实践的结构化概述。如果风险得到妥善管理，组织就能更有信心地采用自主智能体AI。该框架面向希望部署自主智能体AI的组织，无论是内部开发AI智能体还是使用第三方自主智能体解决方案。在之前模型治理框架的基础上，我们在以下四个方面为组织列出了关键考量：

1. 预先评估并限定风险

组织应调整其内部结构和流程，以应对智能体带来的新风险。关键在于首先了解智能体行动所构成的风险，这取决于以下因素：智能体可采取的行动范围、这些行动的可逆性，以及智能体的自主程度。

为了尽早管理这些风险，组织可以在规划阶段通过设计适当的边界来限制智能体的影响范围，例如限制智能体对工具和外部系统的访问。同时还应确保智能体的行动是可追溯和可控的，通过建立健全的智能体身份管理和访问控制来实现。

2. 确保人类切实承担责任

一旦对自主智能体AI的部署"开了绿灯"，组织应采取措施确保人类问责。然而，智能体的自主性可能使传统的、与静态工作流绑定的责任分配变得复杂。多个参与者可能涉及智能体生命周期的不同环节，从而分散了问责。因此，明确界定不同利益相关者的责任至关重要——包括组织内部和与外部供应商之间——同时强调适应性治理，使组织能够快速理解新发展并随着技术演进更新其方法。

具体而言，"人在回路中"（Human-in-the-loop）需要适应以应对自动化偏差（Automation Bias），这一问题随着智能体能力的增强变得更加令人担忧。这包括在智能体工作流中定义需要人类批准的重要检查点，例如高风险或不可逆的操作，以及定期审计人类监督的有效性。

3. 实施技术管控措施与流程

组织应通过在智能体全生命周期实施技术措施，确保AI智能体的安全可靠运行。在开发阶段，应针对规划、工具等新的智能体组件以及仍在成熟中的协议，实施技术管控，以应对这些新攻击面带来的增加风险。

在部署前，应测试智能体的基线安全性和可靠性，包括整体执行准确性、策略遵从性和工具使用等新维度。需要新的测试方法来评估智能体。

在部署期间和之后，由于智能体与其环境动态交互，且并非所有风险都能提前预料，建议逐步推出智能体，并在部署后持续监控。

4. 赋能终端用户的责任意识

智能体的可信部署不仅依赖于开发者，还依赖于终端用户的负责任使用。为实现负责任使用，作为基线，用户应被告知智能体的行动范围、数据访问权限及用户自身的责任。组织应考虑分层提供培训，使员工具备管理人机交互和实施有效监督所需的知识，同时保持其专业技能和基础能力。

本文件是动态的。 我们已与政府机构和领先企业合作汇总了当前最佳实践，但这是一个快速发展的领域，最佳实践也将不断演进。该框架需要持续更新以跟上新发展。我们诚邀各界反馈意见以完善该框架，以及展示如何将该框架应用于负责任的自主智能体部署的案例研究。

---

1 自主智能体AI简介

1.1 什么是自主智能体AI?

自主智能体AI系统（Agentic AI Systems）是指能够通过多步骤规划以实现指定目标的系统，其核心是使用AI智能体。对于什么构成一个智能体尚无共识，但存在一些共同特征——智能体通常具有一定程度的独立规划和行动能力（如搜索网络或创建文件），通过多个步骤来实现用户定义的目标。

在本框架中，我们聚焦于基于语言模型构建的智能体，此类智能体正被越来越广泛地采用。这类智能体使用小型语言模型（SLM）、大型语言模型（LLM）或多模态大型语言模型（MLLM）作为其"大脑"来进行决策和完成任务。不过值得注意的是，软件智能体并非新概念，还存在其他类型的智能体，例如使用确定性规则或其他神经网络来进行决策的智能体。

1.1.1 智能体的核心组件

智能体建立在语言模型之上，从一个简单的基于LLM的应用程序的核心组件开始理解是有帮助的。

1. **模型（Model）**：一个SLM、LLM或MLLM，作为中央推理和规划引擎，即智能体的"大脑"。它处理指令、解释用户输入，并生成上下文适当的响应。
2. **指令（Instructions）**：定义智能体角色、能力和行为约束的自然语言命令，例如LLM的系统提示词（System Prompt）。
3. **记忆（Memory）**：存储并可供LLM访问的信息，可以是短期或长期存储。有时用于让模型从先前的用户交互或外部知识源获取信息。

智能体以与基于LLM的应用程序类似的方式使用模型、指令和记忆。此外，它还具有使其能够完成更复杂任务的其他组件：

1. **规划与推理（Planning and Reasoning）**：模型通常经过训练以进行推理和规划，这意味着它可以输出完成任务所需的一系列步骤。
2. **工具（Tools）**：工具使智能体能够采取行动并与其他系统交互，例如写入文件和数据库、控制设备或执行交易。模型调用工具来完成任务。
3. **协议（Protocols）**：这是智能体与工具及其他智能体之间标准化的通信方式。例如，模型上下文协议（Model Context Protocol, MCP）被开发用于智能体与工具的通信，而Agent2Agent协议（A2A）则定义了智能体之间相互通信的标准。

1.1.2 多智能体架构

在自主智能体系统中，多个智能体协同工作是常见的设置。这有时可以通过让每个智能体专注于特定功能或任务并行工作来提升性能。

多智能体系统的三种常见设计模式包括：

- **顺序式（Sequential）**：智能体在线性工作流中依次工作。每个智能体的输出成为下一个智能体的输入。
- **监督式（Supervisor）**：一个监督智能体协调其下属的专业智能体。
- **群集式（Swarm）**：智能体同时工作，在需要时交接给另一个智能体。

1.1.3 智能体设计如何影响其能力与限制

虽然每个智能体可能具有相同的核心组件，但每个组件的设计可以显著影响智能体的功能。在考虑智能体能做什么时，区分以下两个概念通常是有帮助的：

- **行动空间（Action-space，亦称权限、能力）**：智能体被允许采取的行动范围，由其被允许使用的工具、可执行的交易等决定。
- **自主性（Autonomy，亦称决策能力）**：智能体在多大程度上可以自行决定何时以及如何朝着目标行动，例如定义工作流中要采取的步骤。这可以由其指令和人类参与程度决定。

行动空间

智能体的行动空间主要取决于其可访问的工具，这会影响：

- **可访问的系统**：
- 仅限沙箱：沙箱化的工具（如用于代码执行、数据分析），不能影响任何其他系统
- 内部系统：组织内部的工具，例如能够搜索和更新组织的数据库
- 外部系统：使智能体能够访问外部服务的工具，例如通过第三方预定义API检索和更新数据
- **相对于可访问系统可采取的行动**：
- 读取 vs 写入：智能体可能只能从系统中读取和检索信息，而不能写入或修改系统中的数据。

一种新兴的自主智能体AI模式是**计算机使用智能体（Computer Use Agent）**，其主要工具是对计算机和浏览器的访问权限。这意味着它可以执行人类使用计算机和浏览器能做的任何操作，而无需依赖特定定义的工具和API。这显著增加了智能体可访问和执行的范围。

自主性

智能体的自主性主要取决于其指令组件和人类在系统中的参与程度。

在指令方面，智能体可以被给予不同级别的指令：

- **详细指令和标准操作流程（SOP）**：被指示遵循详细SOP来完成任务的智能体，在每个阶段可做的决策会受到限制。
- **使用自身判断**：被指示使用自身判断来完成任务的智能体，在定义计划和工作流方面拥有更多自由。

另一个相关因素是人类参与程度。在与智能体交互时，人类可以参与到不同程度：

- **智能体建议，人类操作**：人类指导并批准智能体采取的每一步骤。

- **智能体与人类协作**：人类和智能体共同工作。智能体在重要步骤（如写入数据库或执行支付前）需要人类批准。但人类可以随时通过接管智能体的工作或暂停智能体并要求变更来进行干预。
- **智能体操作，人类批准**：智能体仅在关键步骤或失败时需要人类批准，例如删除数据库或执行超过预设金额的支付。
- **智能体操作，人类观察**：智能体在完成任务时不需要人类批准，但其行动可能在事后接受审计。

1.2 自主智能体AI的风险

1.2.1 风险来源

智能体的新组件构成了新的风险来源。风险本身是人们熟悉的——从根本上说，智能体是建立在LLM之上的软件系统。它们继承了传统的软件漏洞（如SQL注入）和LLM特有的风险（如幻觉、偏差、数据泄露和对抗性提示注入）。

然而，风险可以通过不同组件以不同方式表现。例如：

- **规划与推理**：智能体可能产生幻觉并制定错误的任务计划。
- **工具**：智能体可能通过调用不存在的工具、使用错误的输入调用工具，或以有偏差的方式调用工具而产生幻觉。由于工具将智能体连接到外部系统，提示注入或代码注入也可以操纵智能体泄露或以其他方式操纵其可访问的数据。
- **协议**：随着处理智能体通信的新协议的出现，它们也可能被不当部署或遭到破坏，例如部署包含泄露用户数据代码的不受信任的MCP服务器。

当智能体内部的组件或多个智能体之间互动时，风险也可能在系统层面产生。例如：

- **级联效应（Cascading Effect）**：一个智能体的错误可能因其输出传递给其他智能体而迅速升级。例如，在供应链管理中，一个智能体产生幻觉的库存数据可能导致下游智能体重新订购过多或不足的库存。
- **不可预测的结果（Unpredictable Outcomes）**：协同工作的智能体也可能以意想不到的方式竞争或协调。例如，在制造业中，不同的智能体可能参与管理机器和库存。在协调达成生产目标时，智能体可能因复杂的优化算法而产生不可预测的交互，过度或不足地优先分配某一资源或机器，导致意外的瓶颈。

1.2.2 风险类型

由于智能体在现实世界中采取行动，当它们出现故障时，可能导致有害的现实世界影响。组织应意识到以下负面后果：

- **错误行动（Erroneous Actions）**：不正确的行动，例如智能体在错误的日期安排预约或生成有缺陷的代码。具体的有害后果取决于相关行动，如有缺陷的代码可能导致安全漏洞被利用，错误的医疗预约可能影响患者的健康结果。
- **未授权行动（Unauthorised Actions）**：智能体在其允许的范围或权限之外采取的行动，例如根据公司政策或标准操作流程应升级寻求人类批准但未这样做的行动。
- **有偏差或不公平的行动（Biased or Unfair Actions）**：导致不公平结果的行动，特别是在处理不同背景和人口特征的群体时，例如采购中的有偏差的供应商选择、补助金的发放和/或招聘决策。
- **数据泄露（Data Breaches）**：导致敏感数据暴露或被操纵的行动。此类数据可能是个人身份信息或机密信息，如客户详细信息、商业秘密和/或内部通信。这可能是由于安全漏洞——攻击者利用智能体泄露私人信息，或者智能体因未能识别敏感数据而披露了该数据。
- **对关联系统的干扰（Disruption to Connected Systems）**：由于智能体与其他系统交互，当智能体被入侵或出现故障时，可能对关联系统造成干扰，例如删除生产代码库，或对外部系统发送过多请求。

---

2 自主智能体AI模型治理框架

自主智能体AI模型治理框架的四个维度：

1. 预先评估并限定风险
2. 确保人类切实承担责任
3. 实施技术管控措施与流程
4. 赋能终端用户的责任意识

自主智能体AI模型治理框架建立在2020年发布的模型AI治理框架（MGF）中为组织制定的负责任AI实践基础之上，通过突出新兴最佳实践来应对自主智能体AI带来的新问题。这样，组织就可以凭借必要的知识和判断力来开发和使用自主智能体AI。

该框架首先帮助组织**预先评估并限定风险**。它突出了风险评估中应考虑的新风险，以及在规划阶段限制智能体潜在影响范围的设计考量，以及确保智能体可追溯和可控。

虽然智能体可能自主行动，但人类责任仍然适用。一旦对部署自主智能体AI"开了绿灯"，组织应立即采取措施**确保人类切实承担责任**。这包括在组织内外参与智能体生命周期的多个参与者之间明确界定责任；以及采取措施确保"人在回路中"随时间推移保持有效，不受自动化偏差的影响。

为确保智能体的安全可靠运行，组织应在AI全生命周期采用**技术管控措施与流程**。在开发阶段，应实施针对AI智能体新组件（如规划和工具）的护栏。在部署前，应测试智能体的基线安全性和可靠性。在部署后，应随着智能体与环境的动态交互而持续监控。

最后，智能体的可信部署不仅依赖于开发者，还依赖于终端用户。组织有责任通过为用户提供必要信息来**赋能终端用户的责任意识**，使其能够适当使用智能体并实施有效监督，同时保持其专业技能和基础能力。

2.1 预先评估并限定风险

智能体带来了新的风险，特别是在其对敏感数据的访问和通过行动改变环境的能力方面。其自适应、自主和多步骤的特性也增加了意外行动、新兴风险和级联影响的可能性。组织应将这些新维度纳入风险评估，并在早期阶段通过设计适当的边界来限制智能体的影响范围。

在规划使用自主智能体AI时，组织应考虑：

- **确定适合部署智能体的应用场景**，通过考虑可能影响风险可能性和影响的智能体特定因素。
- **通过设计选择预先限定风险**，对智能体访问工具和系统的权限施加限制，并定义健全的身份和权限框架。

2.1.1 确定适合部署智能体的应用场景

风险识别和评估是考虑自主智能体用例是否适合开发或部署的第一步。风险是可能性（风险出现的概率）和影响（风险出现时的严重程度）的函数。

以下非穷尽因素影响自主智能体用例的风险水平：

影响影响程度的因素

因素	描述	示例
	智能体所部署领域和用例对错误的容忍度	执行金融交易（要求高度准确）的智能体 vs 总结内部会议的智能体

因素	描述	示例
智能体部署的领域和用例		
智能体对敏感数据的访问	智能体是否能访问敏感数据，如个人信息或机密数据	需要访问个人客户数据的智能体存在泄露该数据的风险 vs 仅能访问公开信息的智能体
智能体对外部系统的访问	智能体是否能访问外部系统	向第三方API发送数据的智能体可能将数据泄露给第三方，或通过发送过多请求干扰这些系统 vs 仅能访问沙箱或内部工具的智能体
智能体行动的范围	智能体只能读取还是可以修改其可访问的数据和系统	读取 vs 写入：只能从数据库读取的智能体 vs 能向数据库写入的智能体；少量工具 vs 大量工具：只能选择少数预定义工具的智能体 vs 拥有无限浏览器工具访问权限的智能体
智能体行动的可逆性	如果智能体可以修改数据和系统，此类修改是否容易撤销	安排会议的智能体 vs 向外部各方发送电子邮件的智能体

影响可能性的因素

因素	描述	示例
智能体的自主程度	智能体是否能自行定义整个工作流或必须遵循明确定义的流程。更高的自主性可能导致更高的不可预测性，增加错误的可能性。	智能体被提供SOP并被指示在执行任务时遵循 vs 智能体被指示使用自身最佳判断来选择和执行每一步
任务复杂度	任务的复杂程度，与完成任务所需的步骤数量和每一步所需的分析水平相关。更高的复杂度同样增加不可预测性和错误的可能性。	智能体被要求从会议记录中提取关键行动要点 vs 智能体被要求在处理外部信息请求时遵循细致的数据共享政策
智能体对外部系统的访问	智能体是否暴露于外部系统，以及谁维护这些系统。更高的暴露程度使智能体更容易受到提示注入和网络攻击。	智能体只能访问由受信任的内部团队维护的内部知识库 vs 智能体可以访问包含不受信任数据的网络

威胁建模（Threat Modelling） 通过系统地识别攻击者可能采取的特定方式来使风险评估更加严格。常见的自主智能体系统安全威胁包括记忆投毒（Memory Poisoning）、工具滥用和权限破坏。由于自主智能体系统（尤其是多智能体系统）可能变得非常复杂，使用一种称为**污点追踪**

（Taint Tracing） 的方法来映射所有工作流和交互、跟踪不受信任的数据如何在系统中流动通常是有用的。关于如何对自主智能体系统执行威胁建模和污点追踪的更多信息，组织可参考新加坡网络安全局（CSA）的《自主智能体AI安全附录草案》。

> **威胁建模与风险评估的关系**
>
> 威胁建模通过生成具有详细描述的攻击者行动序列、活动和场景的上下文化威胁事件来增强风险评估流程。凭借更相关的威胁事件，风险评估将更加严格和可靠，从而产生更有针对性的控制和有效的分层防御。由于风险评估是持续的，威胁模型应定期更新。
>
> ——改编自CSA《网络威胁建模指南》

2.1.2 通过设计定义智能体的限制和权限来限定风险

在选择了合适的智能体用例后，组织可以通过为每个智能体定义适当的限制和权限策略来进一步限定风险。

智能体限制

组织应考虑在以下方面定义限制：

- **智能体对工具和系统的访问**：定义策略，仅给予智能体完成其任务所需的最小工具和数据访问权限。例如，编码助手可能不需要访问网络搜索工具，特别是如果它已经有策划好的最新软件文档访问权限。
- **智能体的自主性**：对于流程驱动的任务，标准操作流程（SOP）和协议经常被用来提高一致性并减少不可预测性。为智能体工作流定义类似的SOP，约束智能体遵循，而不是让智能体自由定义工作流的每一步。
- **智能体的影响范围**：设计机制和程序，在智能体出现故障时将其下线并限制其潜在影响范围。这可以包括在自包含的环境中运行智能体，限制网络和数据访问，特别是在执行代码执行等高风险任务时。

智能体身份

身份管理和访问控制是组织当前为人类实现可追溯性和问责制的关键手段之一。随着智能体变得更加自主，身份管理也需要扩展到智能体，以跟踪个别智能体的行为并确定谁对每个智能体负有责任。

这是一个不断发展的领域，在健全处理智能体身份方面目前存在差距。例如，当前的授权系统通常具有预定义的静态范围。然而，要在更复杂的场景中安全运行，智能体需要细粒度的权限，这些权限可能根据上下文、风险水平和任务目标动态变化。当前的身份验证系统通常也是基于单一、唯一的个体。此类系统在处理复杂的智能体设置时面临困难，例如智能体代表具有不同权限的多个人类用户行事，或者智能体生成多个子智能体的递归委托场景。

解决方案正在开发中以应对这些问题，例如将OAuth 2.0等成熟标准集成到MCP中。行业也在为智能体开发新的标准和解决方案，如去中心化身份管理和动态访问控制。

在此期间，组织应考虑以下最佳实践以实现智能体的控制和可追溯性：

- **身份识别（Identification）**：智能体应拥有自己的唯一身份，以便向组织、其人类用户或其他智能体表明自身身份。然而，智能体的身份可能需要与监督智能体、人类用户或组织部门挂钩，以实现问责和跟踪。此外，智能体以不同身份行事的情况（例如独立行事或代表指定的人类用户行事）也应予以记录。
- **授权（Authorisation）**：智能体可以根据其角色或手头任务拥有预定义的权限，或者其权限可以由其授权的人类用户动态设置，或两者兼而有之。作为经验法则，人类用户不应能为智能体设置超出该人类用户本身授权范围的权限。此类授权委托应被清晰记录。

> **评估剩余风险**
>
> 剩余风险（Residual Risk）是在缓解措施实施后仍然存在的风险。需要注意的是，即使在努力识别适当的智能体用例并对智能体定义限制之后，始终会存在一定程度的剩余风险，特别是考虑到自主智能体AI发展的速度。最终，组织应评估并确定其自主智能体部署的剩余风险是否处于可容忍的水平并可以接受。

2.2 确保人类切实承担责任

部署智能体的组织和监督智能体的人类对智能体的行为和行动承担责任。但当智能体的行动是从交互中动态和自适应地产生而非基于固定逻辑时，履行这一责任可能充满挑战。多个利益相关者可能参与智能体生命周期的不同环节，分散了问责。最后，自动化偏差——即倾向于过度信任自动化系统，特别是当其过去表现可靠时——随着人类监督日益强大的智能体而变得更加令人担忧。

为应对这些对人类问责的挑战，组织应考虑：

- **在组织内外明确分配责任**，通过在智能体价值链和生命周期中建立问责链，同时强调适应性治理，使组织能够快速理解新发展并更新其方法。

- **确保对智能体的有意义的人类监督的措施**，例如在重要检查点要求人类批准、审计人类批准的有效性，以及用自动化监控来补充这些措施。

2.2.1 在组织内外明确分配责任

作为部署者，组织和人类对智能体的决策和行动承担责任。然而，与AI一样，自主智能体AI的价值链涉及多个参与者。组织应考虑在组织内部以及与价值链中其他组织之间的责任分配。

简化的自主智能体AI价值链：

模型开发者 → 自主智能体AI系统提供商 → 部署组织 → 终端用户

（工具提供商如MCP、API等允许智能体连接到外部系统）

组织内部

在组织内部，组织应在智能体生命周期中为不同团队分配责任。虽然每个组织的结构不同，以下是说明此类责任如何在不同团队之间分配：

关键决策者

- 谁：定义组织战略决策和高级别政策的领导者，如董事会成员、高管、总经理或部门负责人。
- 关键职责可包括：
- 设定使用智能体的高级别目标
- 定义智能体的允许运营用例，包括对智能体数据访问的限制
- 设定整体治理方法，包括风险管理框架和升级流程

产品团队

- 谁：负责将利益相关者需求或业务目标转化为技术自主智能体解决方案的角色，如产品经理、UI/UX设计师、AI工程师、软件工程师。
- 关键职责可包括：
- 定义智能体的设计和需求，以及任何功能控制或分阶段推出方案
- 可靠地实施智能体，即在智能体生命周期中进行开发、部署前测试和部署后监控
- 教育用户负责任地使用智能体产品

网络安全团队

- 谁：负责保护自主智能体系统免受网络威胁的角色，通过实施和管理安全措施、识别漏洞并响应事件，如首席安全官、网络安全专家、渗透测试员。

- 关键职责可包括:
- 定义基线安全护栏和安全设计模板，供技术团队实施或适配到所部署的自主智能体系统
- 定期进行红队测试和威胁建模

用户

- 谁：利用智能体输出来为组织目标做出贡献的任何个人，如做出决策或自动化工作流和实践的公司员工。
- 关键职责可包括:
- 道德和负责任地使用智能体
- 参加所需的培训，遵守使用政策，及时报告智能体的错误或问题

> **发展内部能力以实现适应性治理**
>
> 参与自主智能体AI生命周期的所有团队都应发展理解自主智能体AI的内部能力。由于技术在快速演进，了解新的自主智能体发展的改进和局限——例如计算机使用智能体等新模式或新的智能体评估框架——使组织能够快速调整其治理方法以适应新发展。

组织外部

组织在部署智能体时也可能需要与外部各方合作，如模型开发者、自主智能体AI提供商或外部MCP服务器或工具的托管方。

在这些情况下，组织同样应确保有措施来履行其自身的问责。一些智能体特定的考量包括:

- **在组织与外部各方之间的条款和条件或合同中明确义务的分配**。 特别是，组织应考虑有关安全安排、性能保证或数据保护和保密的条款。如果存在差距，组织应重新评估自主智能体部署是否符合其风险承受能力。
- **维护安全和控制的功能**。 组织应考虑外部各方的产品是否提供让组织维持足够安全或控制水平的功能。这包括强身份验证措施（如限定范围的API密钥、每个智能体的身份令牌）和强大的可观察性（如工具调用和访问历史的日志记录）。如果缺乏此类功能，组织应考虑替代或内部解决方案，或缩小自主智能体用例的范围，例如限制对敏感数据的访问。

终端用户

组织可能向组织内部或外部的用户部署智能体。在此过程中，组织应确保向用户提供充分的信息以便用户对组织进行问责，以及与用户自身责任相关的任何信息。更多信息请参见下文"赋能终端用户的责任意识"。

2.2.2 设计有意义的人工监督

建立有效的人类监督系统

组织应定义需要人类批准的重要检查点或行动边界，特别是在执行敏感行动之前。这可以包括：

- **高风险的行动和决策**，例如编辑敏感数据、高风险领域（如医疗或法律）的最终决策、可能引发法律责任的行动
- **不可逆的行动**，例如永久删除数据、发送通信、执行支付
- **异常或非典型行为**，例如当智能体访问其工作范围之外的系统或数据库时，当智能体选择的配送路线是中位距离两倍长时
- **用户定义的边界**。智能体可能代表具有不同风险偏好的用户行事。除了组织定义的边界外，用户可以被授权定义自己的边界，例如要求批准超过一定金额的购买

除了考虑何时需要批准，组织还应考虑**批准应采取什么形式**。这些考量包括：

- **保持批准请求的上下文性和可消化性**。 在请求人类批准时，保持请求简短明确，而不是提供可能难以解读和理解的长日志或原始数据。
- **考虑所需的人类输入形式**。 对于直接的行动（如访问数据库），人类用户只需批准或拒绝。对于更复杂的情况（如在执行前审查智能体的计划），让人类编辑计划后再给智能体许可可能更有成效。

组织应实施措施以**确保人类监督的持续有效性**，特别是考虑到人类仍然容易受到警报疲劳和自动化偏差的影响。这些措施可以包括：

- **培训人类识别常见的失败模式**，例如不一致的智能体推理、智能体引用过时的政策
- **定期审计人类监督的有效性**

最后，人类监督应辅以**自动化的实时监控**，以升级任何意外或异常行为。这可以通过为某些记录事件实施警报（如未经授权的访问尝试或在指定时间内多次失败的工具调用尝试）、使用数据科学技术识别异常的智能体轨迹，或使用智能体监控其他智能体来实现。更多信息请参见下文"持续测试和监控"。

2.3 实施技术管控措施与流程

自主智能体组件——使智能体区别于简单的基于LLM的应用程序——在实施生命周期的关键阶段需要额外的控制。

组织应考虑：

- **在设计和开发阶段**，设计并实施技术管控。智能体的新组件和能力也需要新的和定制化的控制。根据智能体设计，实施工具护栏和计划反思等控制。此外，通过对工具和数据执行最小权限访问来限制智能体对外部环境的影响。
- **部署前**，测试智能体的安全性和安全性。与所有软件一样，部署前的测试确保系统按预期运行。特别是对于智能体，测试整体任务执行、策略遵从性和工具调用准确性等新维度，并在不同级别和不同数据集上进行测试，以捕获智能体行为的完整范围。
- **部署时**，逐步推出智能体并在生产中持续监控。智能体的自主特性和不断变化的环境使得在部署前测试所有可能的结果变得具有挑战性。因此建议逐步推出智能体，并在部署后进行实时监控以确保智能体安全运行。

2.3.1 在设计和开发阶段使用技术管控

组织应在自主智能体AI系统中设计和实施技术管控以缓解已识别的风险。特别是对于智能体，除了基线软件和LLM控制之外，还应考虑针对以下方面添加控制：

- 新的智能体组件，如规划与推理和工具
- 来自更大攻击面和新协议的增加安全关注

作为示例，以下是一些智能体的示例控制。如需更全面的列表，组织可参考CSA的《自主智能体AI安全附录草案》和GovTech的《智能体风险与能力框架》。

组件	控制措施
规划	· 提示智能体反思其计划是否符合用户指令 · 提示智能体总结其理解并在继续前向用户请求澄清 · 记录智能体的计划和推理供用户评估和验证
工具	· 配置工具要求严格的输入格式 · 应用最小权限原则限制每个智能体可用的工具，通过强大的身份验证和授权来执行 · 对于数据相关工具：不要授予智能体对敏感数据库中表的写入权限（除非严格需要）；配置智能体在输入敏感数据（如密码、API密钥）时让用户接管控制

组件	控制措施
协议	· 在适用时使用标准化协议（如智能体在处理金融交易时使用自主商务协议） · 对于MCP服务器：将受信任的服务器列入白名单，仅允许智能体与白名单上的服务器交互；沙箱化任何代码执行

2.3.2 部署前测试智能体

组织应在部署前测试智能体的安全性。这提供了智能体按预期工作且控制有效的信心。关于软件和LLM测试的最佳实践仍然相关，例如软件系统的单元测试和集成测试，以及为LLM测试选择有代表性的数据集、有用的指标和评估器。组织可参考之前的指导，如《基于LLM的应用安全性和可靠性测试入门套件》。

然而，组织应为智能体**调整其测试方法**。一些考量包括：

- **测试新风险**：除了产生不正确的输出外，智能体可以通过工具采取不安全或意外的行动。组织可以考虑测试：
- **整体任务执行**：智能体是否能准确完成任务
- **策略合规**：智能体是否遵循定义的SOP，并在需要时将操作路由到人类批准
- **工具调用**：智能体是否调用正确的工具、使用正确的权限、正确的输入和正确的顺序
- **鲁棒性**：由于智能体预期对现实世界情况做出反应和适应，测试其对错误和边缘情况的响应
- **测试整个智能体工作流**：智能体可以在无人类参与的情况下按顺序执行多个步骤。因此，除了测试智能体的最终输出外，还应在其整个工作流中进行测试，包括推理和工具调用。
- **单独和协同测试智能体**：除了单个智能体外，还应在多智能体系统层面进行测试，以了解智能体协作时可能出现的新兴风险和行为，如竞争行为或一个智能体被入侵时对其他智能体的影响。
- **在真实或模拟真实的环境中测试**：由于智能体可能需要应对现实世界的情况，测试应在正确配置的执行环境中进行，尽可能模拟生产环境，例如使用工具集成、外部API和行为与部署中一致的沙箱。然而，组织应在真实性需求与过早允许智能体访问影响现实世界的工具的风险之间进行权衡。
- **重复测试并跨多样化数据集**：智能体行为本质上是随机的且依赖上下文。因此，测试应大规模进行并跨多样化数据集，以观察任何意外的低概率行为，特别是如果它们具有高影响。这需要

生成覆盖智能体可能遇到的不同条件的测试数据集，并多次运行这些测试，必要时包括微小扰动。

- **规模化评估测试结果**：可靠地规模化评估测试结果是LLM测试的一个已知挑战。智能体增加了进一步的复杂性，因为其工作流可能很长且包含不能被人类或自动化脚本轻松处理的非结构化信息。组织可以考虑对智能体工作流的不同部分使用不同的评估方法（如对结构化工具调用使用确定性测试，对非结构化智能体推理使用LLM或人类评估）。然而，仍然需要整体评估智能体，以便评估跨步骤的智能体模式。当前行业解决方案因此包括定义LLM或智能体来评估其他智能体。

2.3.3 部署时持续监控和测试

由于智能体是自适应和自主的，组织在部署智能体时应考虑应对意外或新兴风险的机制。

智能体的渐进式部署

组织应考虑**逐步将智能体推出到生产环境**以控制风险暴露量。此类推出可基于以下因素控制：

- **智能体的用户**，例如先向训练有素或经验丰富的用户推出
- **智能体可用的工具和协议**，例如先限制智能体使用更安全的、白名单中的MCP服务器
- **暴露给智能体的系统**，例如先在较低风险的内部系统中使用智能体

持续测试和监控

组织应在部署后**持续监控和记录智能体行为**，并为智能体故障或意外行为建立报告和故障安全机制。这使组织能够：

- **实时干预**：当检测到潜在故障时，停止智能体工作流并升级给人类主管，例如如果智能体尝试未经授权的访问
- **事件发生时进行调试**：记录和追踪智能体工作流的每一步以及智能体之间的交互有助于识别故障点
- **定期审计**：这确保系统按预期运行

监控和可观察性不是新概念，但智能体引入了一些挑战。由于智能体以机器速度执行多个行动，组织面临从监控系统生成的大量日志中提取有意义洞察的问题。当高风险异常被要求实时检测并尽早浮现时，这变得更加困难。

设置监控系统的关键考量包括：

- **记录什么**：组织应确定其监控目标（如实时干预、调试、组件间集成）以识别需要记录的内容。在此过程中，优先监控高风险活动，如更新数据库记录或金融交易。
- **如何有效监控日志**：组织可以考虑以下方法：
- **定义警报阈值**：
  - 程序化、基于阈值的：定义当智能体触发阈值时的警报，例如智能体尝试未经授权的访问或在指定时间范围内进行过多的重复工具调用。
  - 异常值/异常检测：使用数据科学或深度学习技术处理智能体信号并识别可能表明故障的异常行为。
  - 智能体监控其他智能体：设计智能体实时监控其他智能体，标记任何异常或不一致。
- **定义具体干预措施**：对于每种警报类型，考虑干预的级别。应纳入一定程度的人类审查，与风险级别成比例。例如，低优先级警报可在计划时间进行审查，而高优先级警报可能需要暂停智能体执行直到人类审查员可以评估。在发生灾难性智能体故障或被入侵的情况下，应考虑终止和回退解决方案等相称措施。

最后，即使在部署后也应**持续测试自主智能体系统**，以确保其按预期工作且未受到模型漂移或环境其他变化的影响。

2.4 赋能终端用户的责任意识

归根结底，终端用户是使用和依赖智能体的人，人类问责也延伸到这些用户。组织应向终端用户提供充分的信息，以促进信任并实现负责任的使用。

组织应考虑：

- **透明性**：用户应被告知智能体的能力（如智能体对用户数据的访问范围、智能体可以采取的行动）以及当智能体出现故障时用户可以联系的升级渠道。
- **教育**：用户应接受关于正确使用和监督智能体的教育（如应提供关于智能体行动范围、常见失败模式如幻觉、数据使用政策的培训），以及专业技能流失的潜在影响——即随着智能体接管更多功能，基本操作知识可能被侵蚀。因此应提供充分的培训（特别是在智能体普及的领域），以确保人类保持核心技能。

2.4.1 不同用户，不同需求

组织应针对具有不同信息需求的不同用户，使其能够负责任地使用AI。大致而言，终端用户有两种主要原型——与智能体交互的用户，以及将智能体整合到其工作流程中或监督智能体的用户。

- **与智能体交互的用户**（如客服、人力资源智能体——主要面向外部）→ 侧重于透明性
- **将智能体整合到工作流程中的用户**（如编码助手、企业工作流——主要面向内部）→ 在透明性基础上叠加教育和培训

2.4.2 与智能体交互的用户

此类用户通常与代表组织行事的智能体互动，例如客服或销售智能体。这些智能体通常面向外部，但也可以部署在组织内部，例如与组织内其他用户互动的人力资源智能体。

对于这些用户，**侧重于透明性**。组织应分享相关信息以促进信任并便于正确使用智能体。此类信息可以包括：

- **用户的责任**：明确定义用户的责任，例如要求用户核实智能体提供的所有信息。
- **交互**：提前声明用户正在与智能体交互。
- **智能体的行动和决策范围**：告知用户智能体被授权执行的行动和做出的决策范围。
- **数据**：根据组织的数据隐私政策，明确用户数据如何被智能体收集、存储和使用。必要时，在为智能体收集或使用用户数据之前获得用户的明确同意。
- **人类问责和升级**：为用户提供负责智能体的相应人类联络人，用户在智能体出现故障或对决策不满时可以告知。

2.4.3 将智能体整合到工作流程中的用户

此类用户通常将智能体作为其内部工作流的一部分使用，例如编码助手、企业流程自动化。智能体代表用户行事。

对于这些用户，除了上一节中的信息外，**在透明性基础上叠加教育和培训**，使用户能够负责任地使用智能体。关键方面包括教育和培训：

- **智能体的基础知识**
- 相关用例，使用户了解如何将智能体最好地整合到日常工作中，以及应限制使用智能体的场景（如不要将智能体用于机密数据）
- 指导智能体，如提示工程的一般最佳实践、引出特定响应的关键词词汇表

- 智能体的行动范围，使用户了解其能力和潜在影响
- **对智能体的有效监督**
- 常见的智能体失败模式，如幻觉、在错误后陷入循环，使用户能够识别和标记问题
- 持续支持，如定期更新用户最新功能和常见用户错误
- **对专业技能的潜在影响**
- 随着智能体接管入门级任务——这通常是新员工的培训场所——这可能导致用户基本操作知识的流失
- 组织应识别每个职位的核心能力，并提供充分的培训和工作曝光，以确保用户保留基础技能

附录A：更多资源

1. 自主智能体AI简介

什么是自主智能体AI? - AWS, Agentic AI Security Scoping Matrix: A framework for securing autonomous AI systems - WEF, AI Agents in Action: Foundations for Evaluation and Governance - Anthropic, Building effective agents - IBM, The 2026 Guide to AI Agents - McKinsey, What is an AI agent?

自主智能体AI的风险 - GovTech, Agentic Risk & Capability Framework - CSA, Draft Addendum on Securing Agentic AI - OWASP, Multi-Agentic System Threat Modelling Guide - IBM, AI agents: Opportunities, risks, and mitigations - Infosys, Agentic AI risks to the enterprise, and its mitigations

2. 自主智能体AI模型治理框架

预先评估并限定风险

自主智能体治理概述: - EY, Building a risk framework for Agentic AI - McKinsey, Deploying agentic AI with safety and security: A playbook for technology leaders - Bain, Building the Foundation for Agentic AI - OWASP, State of Agentic AI Security and Governance 1.0

风险评估与威胁建模： - OWASP, Agentic AI – Threats & Mitigations - OWASP, Multi-Agentic System Threat Modelling Guide - Cloud Security Alliance, Agentic AI: Understanding Its Evolution, Risks, and Security Challenges - EY, Building a risk framework for Agentic AI

智能体限制与智能体身份： - Meta, Agents Rule of Two: A Practical Approach to AI Agent Security - OpenID, Identity Management for Agentic AI

确保人类切实承担责任

组织内外的责任分配： - Carnegie Mellon University, The 'Who', 'What', and 'How' of Responsible AI Governance - CSA and FAR.AI, Securing Agentic AI: A Discussion Paper - McKinsey, Accountability by design in the agentic organization

设计有意义的人工监督： - Partnership on AI, Prioritizing real-time failure detection in AI agents - Permit.IO, Human-in-the-Loop for AI Agents: Best Practices, Frameworks, Use Cases, and Demo

实施技术管控措施与流程

技术管控： - GovTech, Agentic Risk & Capability Framework - CSA, Draft Addendum on Securing Agentic AI

测试与评估： - Microsoft, Microsoft Agent Evaluators - AWS, AWS Agent Evaluation - Anthropic, Demystifying evals for AI agents - IBM, What is AI Agent Evaluation?

监控与可观察性： - Microsoft, Top 5 agent observability best practices for reliable AI

赋能终端用户的责任意识 - Zendesk, What is AI transparency? A comprehensive guide - HR Brew, Salesforce's head of talent growth and development shares how the tech giant is training its 72,000 employees on agentic AI - Harvard Business Review, The Perils of Using AI to Replace Entry-Level Jobs

---

附录B：征集反馈意见和案例研究

征集反馈意见： 本文件是动态的，我们诚邀各界就如何更新或完善该框架提出建议。以下问题可作为指导：

- **自主智能体AI简介**：对自主智能体AI系统的描述是否准确且足够全面，使读者能够清晰了解自主智能体AI的治理挑战？是否有其他风险应被纳入？
- **拟议的模型治理框架**：该框架的四个维度是否具有实际性和适用性？是否应纳入其他维度？对于每个维度，是否有应纳入的特定治理和技术挑战及最佳实践？

征集案例研究： 我们也邀请各组织提交自身的自主智能体治理经验作为案例研究，展示框架特定方面如何实施，以作为其他组织可参考的负责任部署的实践范例。案例研究理想情况下应涉及组织部署的自主智能体用例，展示框架的某一维度。虽然不穷尽列举，但我们特别关注以下方面的良好实践案例研究：

维度	示例案例研究
预先评估并限定风险	· 定义用例以降低风险但最大化智能体效益 · 通过定义SOP和工作流限制智能体自主性 · 限制智能体对工具和系统的访问 · 如何为智能体实施身份管理，以及其与组织中人类身份的交互
确保人类切实承担责任	· 在组织内分配自主智能体部署的责任 · 评估自主智能体用例中何时需要人类批准，以及如何实施此类批准请求
实施技术管控措施与流程	· 为智能体设计和实施技术管控 · 如何进行智能体安全测试 · 如何设置监控和可观察性机制，包括定义警报阈值和处理大量智能体相关数据
赋能终端用户的责任意识	· 向与智能体交互和使用智能体的内外部利益相关者提供信息 · 培训人类监督者实施有效监督

有关案例研究可能的样式，请参考我们之前的《AI模型治理框架》中的案例研究。

请注意，任何反馈意见和案例研究可能被纳入该框架的更新版本，贡献者将被相应致谢。请通过以下链接提交您的反馈意见和案例研究：https://go.gov.sg/mgfagentic-feedback

版本1.0 | 2026年1月22日发布