

In collaboration
with Capgemini



AI Agents in Action: Foundations for Evaluation and Governance

WHITE PAPER
NOVEMBER 2025



Contents

Foreword	4
Executive summary	5
Introduction	6
1 Evolving technical foundations of AI agents	8
1.1 The software architecture of an AI agent	8
1.2 Communication protocols and interoperability	10
1.3 Cybersecurity considerations	12
2 Foundations for AI agent evaluation and governance	13
2.1 Classification	14
2.2 Evaluation	19
2.3 Risk assessment	22
2.4 Governance considerations for AI agents: a progressive approach	25
3 Looking ahead: multi-agent ecosystems	29
Conclusion	30
Contributors	31
Endnotes	34

Disclaimer

This document is published by the World Economic Forum as a contribution to a project, insight area or interaction. The findings, interpretations and conclusions expressed herein are a result of a collaborative process facilitated and endorsed by the World Economic Forum but whose results do not necessarily represent the views of the World Economic Forum, nor the entirety of its Members, Partners or other stakeholders.

© 2025 World Economic Forum. All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, including photocopying and recording, or by any information storage and retrieval system.

Foreword



Roshan Gya
Chief Executive Officer,
Capgemini Invent



Cathy Li
Head, Centre for AI
Excellence, Member of
the Executive Committee,
World Economic Forum

In recent years, organizations have moved beyond predictive models and chat interfaces to experiment with artificial intelligence (AI) in more transformative ways. AI agents are now emerging as integrated collaborators in business, public services and everyday life. The adoption of AI agents could bring significant gains in efficiency, altered kinds of human-machine interaction and the advent of novel digital ecosystems.

This transition faces multiple obstacles that need to be addressed. Moving from models to agents represents more than a technical milestone and requires organizations to rethink how they design, evaluate and govern advanced agentic systems. Many companies are now questioning what agents can accomplish alongside the practical steps needed to adopt and deploy them safely, responsibly and effectively.

This paper was developed to help answer those questions. By mapping the evolving foundations of agentic systems, classifying their roles, identifying new ways to evaluate them and outlining progressive governance approaches, the paper offers practical guidance for leaders navigating adoption in real-world contexts.

Through the AI Governance Alliance, the World Economic Forum and Capgemini are advancing this subject in collaboration with the AI community, signalling that now is the time to prepare for an agentic future. If adopters start small, iterate carefully and apply proportionate safeguards, agents can be deployed in ways that amplify human capabilities, unlock productivity and establish a foundation for more complex multi-agent ecosystems to emerge over time. Unless a careful and deliberate approach to adoption is adopted, untested use cases could outpace oversight and lead to misaligned incentives, emergent risks and loss of public trust.

As with any transformative technology, the opportunities presented by AI agents must be accompanied by a responsibility to guide their development and deployment with care. Through cross-functional efforts and collaborative governance, AI agents can be integrated in ways that amplify human ingenuity, promote innovation and improve overall quality of life. This paper is a step in that direction, offering guidance to help early adopters navigate the complex and often uneven path of AI agent adoption.

Executive summary

This paper explores the emergence of AI agents, outlining their technical foundations, classification, evaluation and governance to support safe and effective adoption.

This report has been tailored mainly for adopters of AI agents, including decision-makers, technical leaders and practitioners seeking to integrate AI agents into organizational workflows and services.

While AI agents are gaining traction, there remains limited guidance on how to design, test and oversee them responsibly. This paper aims to help fill that gap by providing a structured foundation for the safe and effective deployment of these systems.

The paper makes three key contributions. Firstly, it covers the technical foundations of AI agents, including their architectures, protocols and security considerations. Secondly, it offers a functional

classification that differentiates agents by their role, autonomy, authority, predictability and operational context. Thirdly, it suggests a progressive governance approach that directly connects evaluation and safeguards to an agent's task scope and deployment environment.

Together, these elements guide adopters with a conceptual blueprint for moving from experimentation to deployment. The report highlights the importance of aligning adoption with evaluation and governance practices to ensure that AI agents are successfully deployed while trust, safety and accountability are maintained.

Introduction

AI agents are shifting from prototypes to deployment, bringing both transformative opportunities and novel governance challenges.

AI agents are gradually becoming embedded in an increasing number of tasks, workflows and use cases that span cloud and edge computing, leading the way to more widespread adoption. As the transition from prototyping to deployment accelerates, current adoption remains concentrated among early adopters. According to a recent global survey of executives, 82% of organizations plan to integrate agents within the next one to three years, indicating that most efforts are still in the planning or pilot phase,¹ while moving towards wider adoption.

The concept of software agents has been studied for decades in fields such as robotics, autonomous systems and distributed computing. What is different

today is the rise of data-driven models, particularly generative artificial intelligence (AI) and large language models (LLMs), which are enabling the emergence of a new generation of LLM-based agents. These systems can generate plans, simulate reasoning and adapt their behaviour through feedback mechanisms in ways that were previously not possible. This evolution has sparked a new wave of experimentation, with researchers and companies rapidly creating prototypes of agents in various fields. This report focuses mainly on LLM-based agents (“AI agents” is sometimes used in short), whose growing capabilities create both significant opportunities for adoption and a new set of challenges in governance and safety.

FIGURE 1

Foundations for the responsible adoption of AI agents





LLM-based AI agents, for example, introduce new risks such as goal misalignment, behavioural drift, tool misuse and emergent coordination failures that traditional software governance models are unable to manage. Unlike conventional software, agents are increasingly assuming roles that resemble those of human decision-makers rather than static tools. This means that governance models designed solely for access control and system reliability are no longer sufficient. A more useful comparison is the governance applied to human users, who must earn permissions, accountability and trust by demonstrating performance over time. Similarly, trust in AI agents can be established by testing their behaviour against validated cases, running them in human-in-the-loop configurations and gradually expanding autonomy only once reliability has been sufficiently demonstrated. In both cases, the principle of least privilege remains essential, with access limited to information and actions necessary for the task.

This report aims to provide a forward-looking analysis of the evolving landscape of AI agents, focusing on the capabilities, infrastructure, classification and safeguards necessary for responsible deployment. To this end, it is structured around four foundational

pillars across classification, evaluation, risk assessment and governance, which together form the foundation for a progressive approach to adoption and deployment. Figure 1 presents the general content of this report, which helps guide the responsible adoption and deployment of AI agents.

The goal is to equip adopters, providers, technical leaders, organizational decision-makers and other stakeholders with a shared understanding of the current state of agentic systems and emerging oversight practices. Building on established AI governance principles and frameworks, such as those developed by the Organisation for Economic Co-operation and Development (OECD),² National Institute of Standards and Technology (NIST),³ International Organization for Standardization (ISO)/International Electrotechnical Commission (IEC)⁴ and others, this paper introduces additional principles addressing autonomy, authority, operational context and systemic risk that extend existing governance guidance from an agent-focused lens. The insights have been informed by working group meetings, workshops and extensive interviews with members of the Safe Systems and Technologies working group of the AI Governance Alliance.

Evolving technical foundations of AI agents

The architecture, protocols and security models of AI agents dictate how they integrate into organizations and interact with the world.

While the core architecture of AI agents is beginning to take shape, practices for agent deployment, integration and governance remain nascent. As organizations begin to “hire” AI agents to support or augment human teams, or perform tasks that impact the physical world, adoption should be treated

with the same level of rigour as onboarding a new employee, including clearly defined roles, safeguards and structured oversight mechanisms. This section outlines the technical foundations that enable agentic systems and the architecture decisions that shape how they are built, deployed and governed.

1.1 The software architecture of an AI agent

“Building agents requires not just engineering but also orchestration and coordination between models, tools, data sources and humans.”

The adoption of LLM-based agents by industry marks a broader shift in software development from rigid, rules-based systems to more flexible, intent-driven interactions. For instance, in call centres, early chatbots that followed scripted decision trees are now giving way to agentic systems capable of understanding intent, managing context, and escalating decisions more dynamically. This evolution towards agentic AI represents a fundamental change in control and autonomy, where tasks traditionally performed by humans are delegated to machines.

To enable this shift, AI agents draw on four technological paradigms:

- **Classical software:** deterministic logic and rule-based execution
- **Neural networks:** pattern recognition and statistical learning
- **Foundation models:** general-purpose, adaptive systems that interpret instructions and act contextually
- **Autonomous control:** mechanisms that enable systems to plan, coordinate and act with minimal human oversight

As a result, building agents requires not just engineering but also orchestration and coordination

between models, tools, data sources and humans. This layered setup introduces new complexity in how agents behave, generalize and interact with their environment, reinforcing the need for structured scaffolding.

Today, AI agent architectures are organized into three interconnected layers, consisting of **application, orchestration and reasoning**, which collectively enable intelligent, context-aware and business-aligned automation. At a high level, agent architectures are designed to interface with users and systems, coordinate complex tasks using external tools and application programming interfaces (APIs), and support decision-making through a combination of language models, reasoning modules and control logic. Together, these layers provide the technical foundation that underpins how agents operate.

The **application layer**, along with protocols such as Model Context Protocol (MCP) and agent-to-agent protocol (A2A), integrates the agent into specific processes or user workflows. It receives input through user interfaces or APIs and translates it into structured signals. Application logic applies domain-specific rules and constraints to ensure the agent’s output (i.e. forecast, decisions, actions, messages, etc.) is aligned with user expectations and business requirements. This layer can run in the cloud or on-prem in edge computing equipment.

“Understanding this architecture is key to anticipating how agents will engage with users and systems, coordinate workflows and make context-aware decisions.”

The **orchestration layer** (framework layer) governs how the agent interprets inputs, invokes tools and coordinates tasks. While some LLM providers⁵ have integrated tools directly into their solutions, this can create rigid and vendor-locked systems. Agentic frameworks overcome this by standardizing tools and systems integration, remaining LLM-agnostic and spanning multiple workloads across cloud and edge. This enables AI agents to employ a range of reasoning strategies and support features, such as code execution or search, and use protocols like MCP to connect with enterprise resources, including databases and customer relationship management (CRM) systems. Most agents also include specialized sub-agents that handle distinct tasks, which makes them functionally part of a multi-agent system. The orchestration layer is critical in this regard, as it coordinates sub-agents, assigns responsibilities and manages dependencies between them. It also enables model switching, allowing organizations to assign different models to various tasks based on their complexity, cost or performance. Importantly, AI agents have a unique architecture that can be extended beyond the organization’s security perimeter. Their ability to invoke external tools and communicate with other agents enables them to

operate beyond traditional network boundaries, introducing novel cybersecurity concerns.

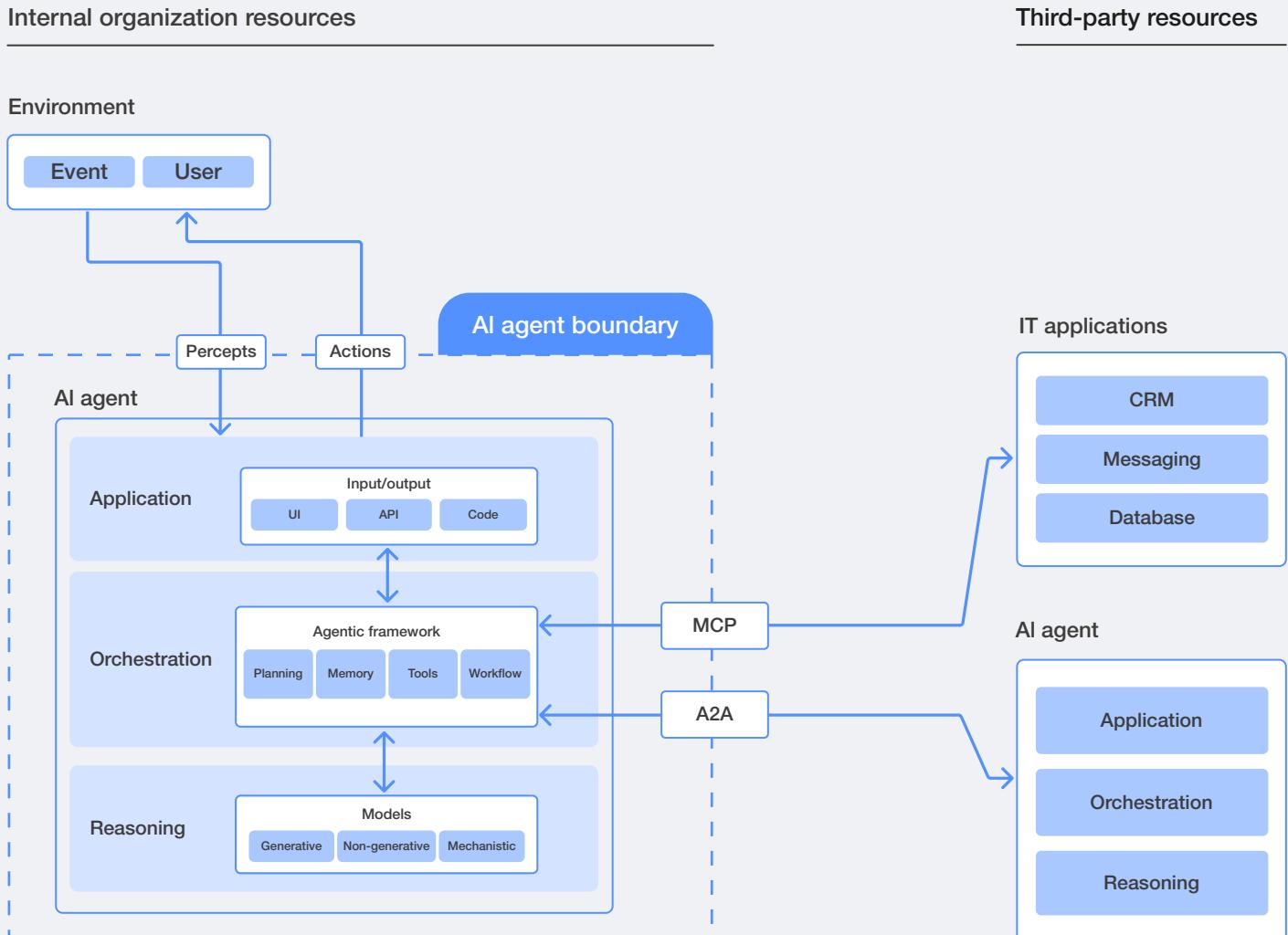
The **reasoning layer** underpins the agent’s ability to generate, predict, classify or apply rules in pursuit of its goals. Depending on the task, the reasoning layer can draw on a range of models, including deterministic, rule-based approaches and classical machine learning, as well as small or large language models and other generative architectures. The choice of model shapes how the agent processes information, adapts to context and ultimately carries out its assigned role.

Figure 2 illustrates this layered architecture, showing how internal components across application, orchestration and reasoning work together to support dynamic agent behaviour while maintaining secure boundaries across organizational systems.

In combination, these layers constitute the technical backbone that governs agent functionality. For organizations implementing AI agents, understanding this architecture is key to anticipating how agents will engage with users and systems, coordinate workflows and make context-aware decisions.

FIGURE 2

Software architecture of an AI agent



1.2 Communication protocols and interoperability

 MCP has gained widespread support across leading agent frameworks and is increasingly viewed as a core mechanism.

The landscape of advanced LLM-based agents is supported by new protocols that enable more seamless integration and collaboration. The MCP, for example, aims to standardize the connection between enterprise software systems, external data sources and agents, while protocols such as A2A and the AGNTCY architecture's agent connect protocol (ACP) offer tools to facilitate interaction between varying AI agents, forming the interoperability layer for multi-agent systems (MAS). As these protocols are implemented across cloud platforms, enterprise networks and edge devices, they are necessary for running agentic code while connecting with real-world sensor data and systems.

Introduced by Anthropic in late 2024, MCP⁶ enables agents to connect with internal or

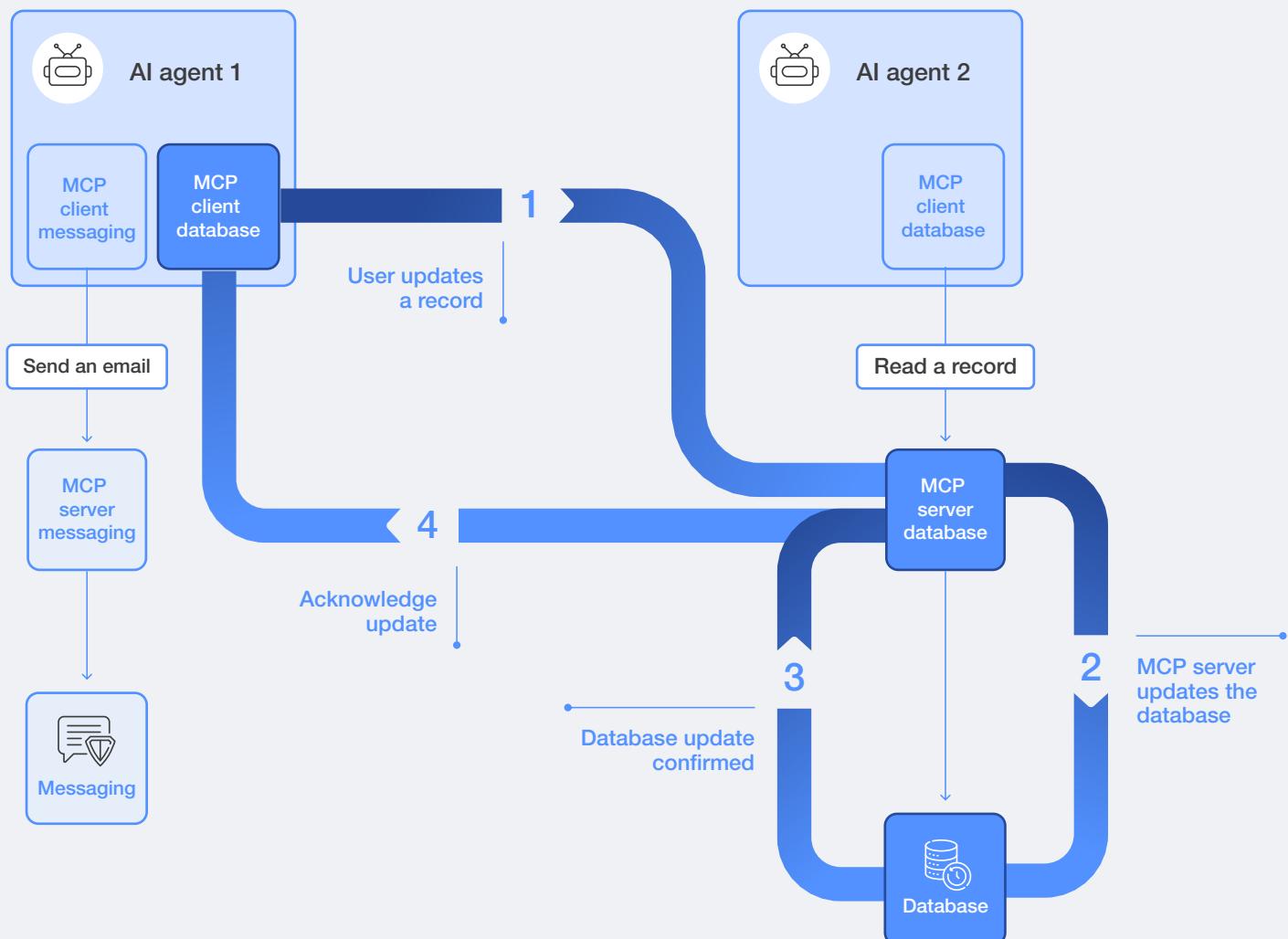
external data sources, APIs and enterprise systems through a standardized protocol. Rather than developing bespoke integrations for each agent-task pairing, MCP allows agents to act as clients that request access to services via MCP-compliant servers. For example, an agent using MCP can check a calendar, retrieve emails, update database content or update CRM records through a shared interface. This significantly reduces friction, speeds up deployment and supports modular plug-and-play capabilities across tools and environments.

MCP has gained widespread support across leading agent frameworks and is increasingly viewed as a core mechanism for connecting agents to the broader enterprise infrastructure.

FIGURE 3

Illustration of MCP-based agent communication

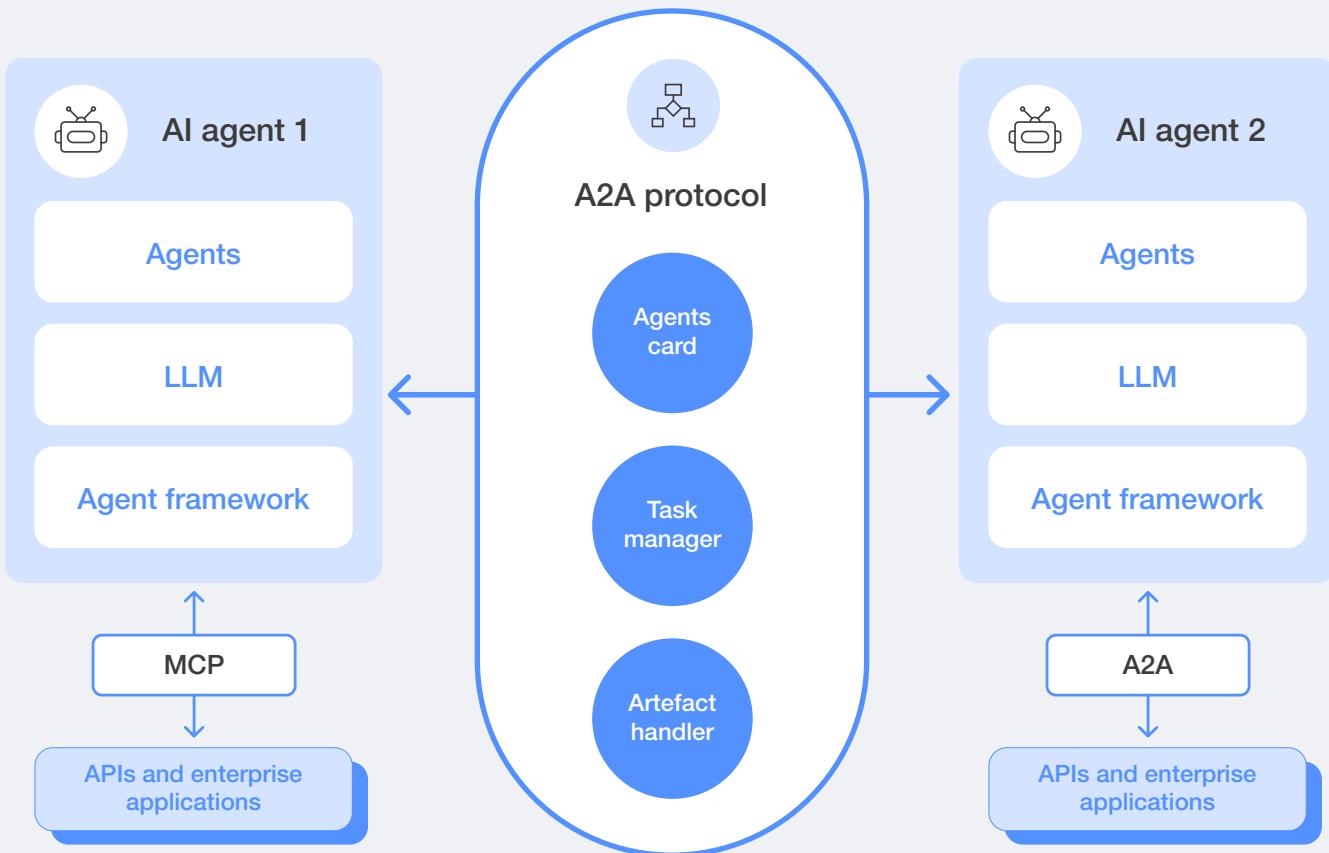
Overview of MCP



Where MCP focuses on communication between agents and external or internal systems, protocols like A2A enable agents to discover each other, interact, collaborate and delegate tasks, whether operating within an organization's security perimeter or outside it. These protocols address a growing need in complex environments where multiple agents work together across organizational or technical boundaries, enabling agents from different vendors to communicate effectively.

Released by Google in April 2025,⁷ A2A operates through a common communication interface and introduces the concept of agent cards (similar to model cards⁸), which are structured descriptions of an agent's identity, along with its capabilities and skills. This allows for automatic discovery and coordination between agents and systems.

FIGURE 4 Illustration of agent-to-agent communication protocol



Beyond communication and discovery, new standards are also emerging that address how agents transact and exchange value. Released by Google in September 2025, the Agent Payments Protocol (AP2)⁹ enables secure, auditable transactions under user-defined limits. Unlike MCP and A2A, which focus on data exchange and task coordination, AP2 addresses complex financial operations.

Despite this progress, interoperability remains a key challenge. Technical compatibility alone does not guarantee successful coordination

between agents. Strategy, privacy and security considerations often shape how and whether systems should be integrated and are important for enterprises to carefully consider.

For example, communication between different agents could raise concerns about access control, data confidentiality or compliance across jurisdictions. Choosing whether to expose a capability to other agents becomes a governance decision as much as a technical choice.

1.3 Cybersecurity considerations

❸ Security strategies have evolved from perimeter defences to layered “defence in depth” and more recently to the zero-trust model.

As AI agents move into enterprise and consumer-facing environments, they extend rather than replace existing security challenges. Security strategies have evolved from perimeter defences to layered “defence in depth,” and more recently to the zero-trust model.¹⁰ These changes reflect broader transformations such as cloud adoption, distributed workforces and interconnected ecosystems, all of which have already weakened the notion of a clear boundary between internal and external networks. Agents build on this trajectory but add additional layers of risk that must be managed proactively.

By autonomously invoking tools and communicating across organizational lines (e.g. via MCP and A2A), agents embed external services, databases and peer agents into enterprise workflows. This multiplication of identities and connections makes identity management, micro-segmentation and ongoing verification of agent activity essential.

While protocols such as MCP and A2A can streamline integration, they also expand the attack surface¹¹ by introducing new external dependencies and interfaces, as illustrated in Figure 2. The very interoperability that enhances agent capabilities also exposes enterprises to unpredictable inputs and vulnerabilities from third parties. For adopters, this means that every agent interaction should be treated as untrusted by default, and that verifying identity, permissions and context is necessary before granting access.

Finally, agents can be misused.¹² They might be exploited through design flaws or prompt injections, or even intentionally deployed for malicious purposes, such as accessing private data or spreading misinformation. Unlike traditional attacks, autonomous agents can act with speed and persistence, making attribution and accountability harder. Organizations should prepare for this by implementing strong audit trails, incident response plans and clear accountability structures.



Foundations for AI agent evaluation and governance

A structured foundation for evaluating and governing AI agents enables consistent assessment and oversight across contexts.

 **Systematic classification is important because it provides a common basis for comparing agents, anticipating risks and linking evaluation and governance.**

As AI agents mature and adoption increases, a functional understanding of their roles and properties is beginning to take shape. Rather than classifying agents solely by modality (e.g. text, speech, vision) or domain (e.g. customer service, decision support, workflow orchestration), it is more effective to evaluate them according to their intended purpose, core properties and operating context. This approach creates a clearer foundation for assessing impacts and designing safeguards that are proportionate to an agent's role. Systematic classification is important because it provides a common basis for comparing agents, anticipating risks and linking evaluation and governance decisions to the realities of how an agent operates. Without it, oversight risks may become inconsistent, reactive or disconnected from an agent's actual capabilities and environment.

To establish this foundation, this report introduces four foundational pillars which, in combination, provide a structured approach to assessment and adoption:

- **Classification:** Establish the agent's characteristics and operational context to inform downstream assessment.
- **Evaluation:** Generate evidence of performance and limitations in representative settings.
- **Risk assessment:** Analyse potential harm using classification and evaluation as inputs.
- **Governance:** Translate classification, evaluation and risk assessment results into safeguards and accountability proportionate to the agent's profile.

These foundations apply to diverse AI agents, encompassing both virtual and embodied systems in different operational contexts. They provide a consistent basis for assessing

performance, identifying risks and establishing governance mechanisms that scale with an agent's autonomy, authority and function.

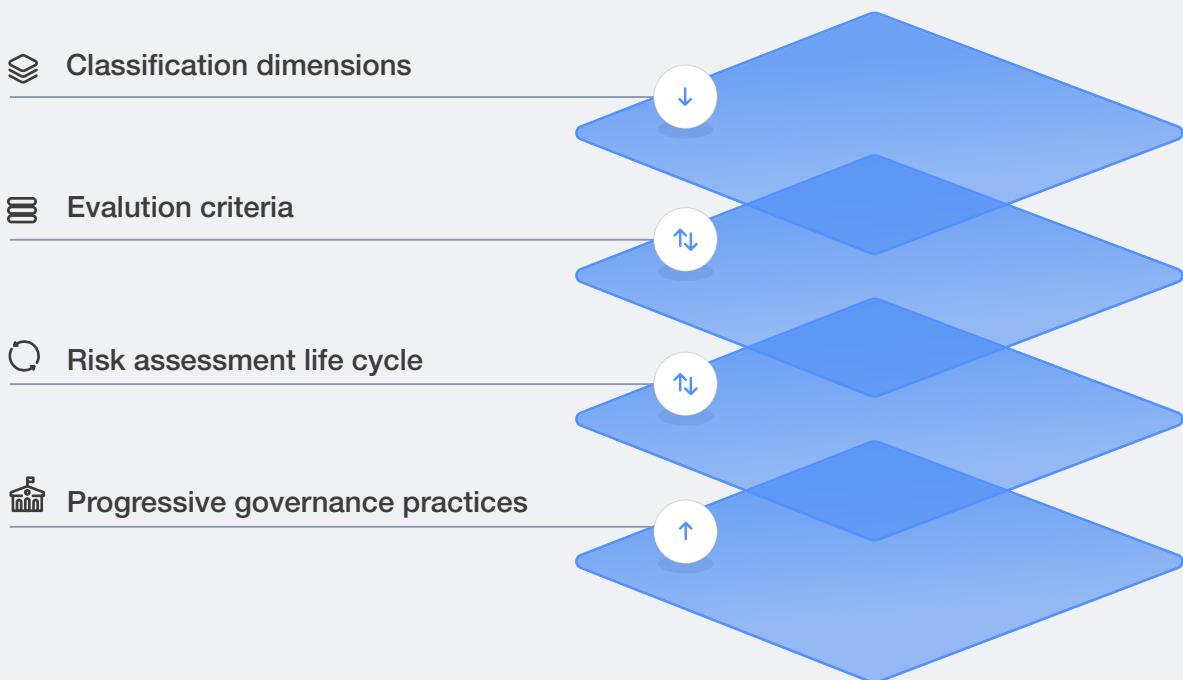
To address classification, evaluation, risk assessment and governance, it is useful to distinguish between two main stakeholder perspectives:¹³

- **Provider:** Refers to organizations or individuals that supply AI systems, platforms or tools. Their responsibilities include ensuring that products are developed and maintained in accordance with responsible and ethical guidelines, and that the necessary documentation and support are provided.
- **Adopter:** Refers to individuals within an organization who use AI systems, encompassing responsibilities such as procurement and deployment. Procurement involves the responsibility of acquiring AI solutions for organizational use by conducting due diligence and ensuring that all AI agent solutions comply with organizational policies and regulatory requirements. Deployment is the responsibility for implementing AI systems in accordance with documented requirements and plans, while ensuring that risks and impacts of the AI agent are properly assessed and managed.

The adopter depends on the provider for transparent documentation, model and system specifications, and sufficient performance and risk information to support responsible deployment and oversight throughout the system life cycle.

The four pillars form a continuous and parallel progression in which classification provides structure, evaluation establishes evidence, risk assessment identifies and mitigates potential harms, and governance translates those insights into safeguards and accountability.

FIGURE 5 | Foundations for AI agent evaluation and governance



2.1 Classification

Classification defines an agent's characteristics and operating context to guide evaluation, risk assessment and governance.

To support evaluation and risk assessment, agents can be described across a set of dimensions that capture both their internal characteristics and the external contexts in which they operate. These dimensions provide a structured approach to analyse and compare agents across applications, ensuring clarity about their design choices and real-world effects.

In combination, the proposed dimensions define how an agent operates, what actions it is permitted to take and the complexity of the context it is deployed in. The agent's overall impact can be seen as a profile that emerges from the interaction of these dimensions, reflecting the benefits or risks of its application in practice.¹⁴

Function refers to the specific role, purpose or set of tasks the agent is designed to perform. It describes what the agent does in practice, independent of the environment it is deployed in. For example, a coding co-pilot that generates software snippets and a triage assistant that prioritizes patients in an emergency department have distinct functions, even though both operate in digital workflows.

Role reflects the breadth of tasks an agent can perform. Specialized agents are narrowly focused and optimized for specific domains, while generalized agents can adapt across domains to address a broader range of tasks or challenges. For instance, a tax-filing agent designed only to prepare returns is specialized, whereas a personal digital assistant that manages scheduling, email drafting and online search operates as a generalist agent.

Predictability describes the stability and repeatability of agent behaviour. Deterministic agents produce consistent, identical outputs when given the same inputs, which makes their performance highly predictable and easier to validate. Non-deterministic agents, by contrast, may evolve, learn or generate variable outputs over time.¹⁵ This variability can support creativity, adaptation and exploration, but it reduces the reliability of producing identical results under identical conditions. For adopters, predictability determines how much confidence they can place in an agent's outputs, how reproducible those outputs are, and what level of oversight is required to manage variability in practice.

Autonomy captures the degree to which an agent can define and pursue objectives. The spectrum ranges from simple command-response systems to

Establishing levels of autonomy can help organizations set clear expectations for functionality and implement proportionate governance mechanisms.

agents capable of planning and executing actions independently across authorized environments. Autonomy in this context refers to an agent's capacity to decide when and how to act toward a goal, adapting to changing conditions without human guidance. Automation, on the other hand, refers to systems that execute predefined functions reliably under specified conditions without human intervention. The key distinction is that autonomy entails decision-making flexibility (i.e. choosing what to do), whereas automation emphasizes execution reliability (i.e. doing what the system is programmed to do).

In the automotive sector, SAE International's *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*¹⁶ framework defines driving automation from Level 0 (no automation) to Level 5 (full automation). A similar spectrum can be applied to AI agents. This spectrum can be conceived of as moving from no autonomy (for example, a simple chatbot that only answers user queries) to full autonomy (for example, a customer service agent that automates interactions, resolves queries and personalizes responses using a company's knowledge base). Establishing levels of autonomy can help organizations set clear expectations for functionality and implement proportionate governance mechanisms.

Authority defines the actions an agent is permitted to take. It sets the boundaries of system access, such as permissions to use tools, interact with databases or execute transactions. Like autonomy, authority exists on a sliding scale, from read-only access to full administrative control.

Autonomy and authority can be combined in different ways, depending on an agent's purpose and design. They are not inherent system properties but design choices that can be made based on the agents' intended functions, risk considerations and oversight requirements. They can also be calibrated during assessment or adjusted in real time.

Operational context refers to the use case and environment in which the agent operates. The environment is especially critical, as it determines observability, predictability of outcomes, interaction with other agents and how conditions evolve over time.¹⁷

Use case defines the domain and environment where the agent performs its distinct function for stakeholders. For example, an autonomous cleaning agent in the residential sector performs household vacuuming and floor cleaning as part of routine home maintenance.

Environment represents the operating conditions the agent functions under, ranging from simple and predictable settings to complex, uncertain and dynamic contexts. A complex environment is one where the agent navigates and acts under uncertainty, with incomplete or noisy information, unpredictable outcomes, changing conditions over time, continuous ranges of possible actions or states, and interactions with other agents whose behaviour also affects results. By contrast, a simple environment is one where the agent operates with complete information, predictable and static outcomes, independent episodes, a finite set of states or actions, and no need to consider other actors.

FIGURE 6

Classification dimensions

Agent characteristics

1. Function

What does the agent do?

2. Role

Specialist ————— Generalist

3. Predictability

Deterministic ————— Non-deterministic

4. Autonomy

Low ————— High

5. Authority

Low ————— High

Operational context

6. Use case

Application domain and environment where the agent performs its function

7. Environment

Simple ————— Complex



For organizations adopting AI, understanding and clearly defining the operational context is essential to ensure effectiveness in actual deployment.

An example of an operational context could be a fraud detection agent in online banking. The agent accesses transaction data and user history, but cannot fully observe external factors like user intentions or hidden fraud tactics. It functions stochastically, with outcomes influenced by unpredictable variables such as varying fraud methods or user behaviours, rather than guaranteed results. The setup is sequential, refining risk assessments with each detection. Operating in a fast-changing environment, it requires continuous monitoring by human reviewers and other security systems.

For organizations adopting AI, understanding and clearly defining the operational context is essential to ensure effectiveness in actual deployment settings. Potential issues can be

mitigated by adjusting agent parameters, like autonomy and authority, and/or by constraining the context in which the agent operates. Examples include limiting a robot to a controlled zone or confining a software agent to a sandbox.

An AI agent's role, autonomy, authority, predictability and operational context collectively shape its overall impact, defined as the degree of benefit or harm it may generate. Highly autonomous, authorized and non-deterministic behaviour in a complex operational context may deliver strong performance but also carry greater risks.

The following example illustrates how these dimensions can be applied in practice through the classification of a basic AI agent, a robot vacuum cleaner.



CASE STUDY 1

Robot vacuum cleaner – classification

Robot vacuum cleaner

Agent characteristics

1. Function

Autonomous indoor navigation and cleaning floors

2. Role

Specialist | Generalist

3. Predictability

Deterministic | Non-deterministic

4. Autonomy

Low | High

5. Authority

Low | High

Operational context

6. Use case

A home vacuum robot operates in the household services domain, autonomously navigating a residential environment to clean floors for occupants.

7. Environment

Simple | Complex

Robot vacuum cleaner – classification

- Function:** The primary function is autonomous indoor navigation and cleaning, interpreting spatial layouts, avoiding obstacles and adapting to changing floor conditions. While it operates autonomously within mapped areas and schedules, it does not make decisions that affect other systems or safety-critical outcomes.
- Role:** Specialist – it only does one specific job, vacuuming the floors.
- Predictability:** Deterministic – it follows specific instructions and task planning but may follow unspecified routes.

- Autonomy:** Medium – it operates independently within mapped areas of the home's floorplan.
- Authority:** Low – it is limited to sensing, movement and vacuuming.

Operational context

Use case: Home vacuuming

Environment: Moderate – the environment is primarily household environments with occasional dynamic obstacles.

As agents become more embedded in tools, platforms and workflows, the proposed dimensions can help organizations define specific agent roles and levels of integration while evaluating benefits and limitations in context and implement oversight mechanisms that match their capabilities. Taking these dimensions into consideration can help providers and adopters to:

- **Clarify functional scope:** Define what an agent is designed to do, under what conditions and where its responsibilities begin and end.
- **Support assessment:** Evaluate the technical, organizational, safety and security implications of deploying specific agents in their contexts.

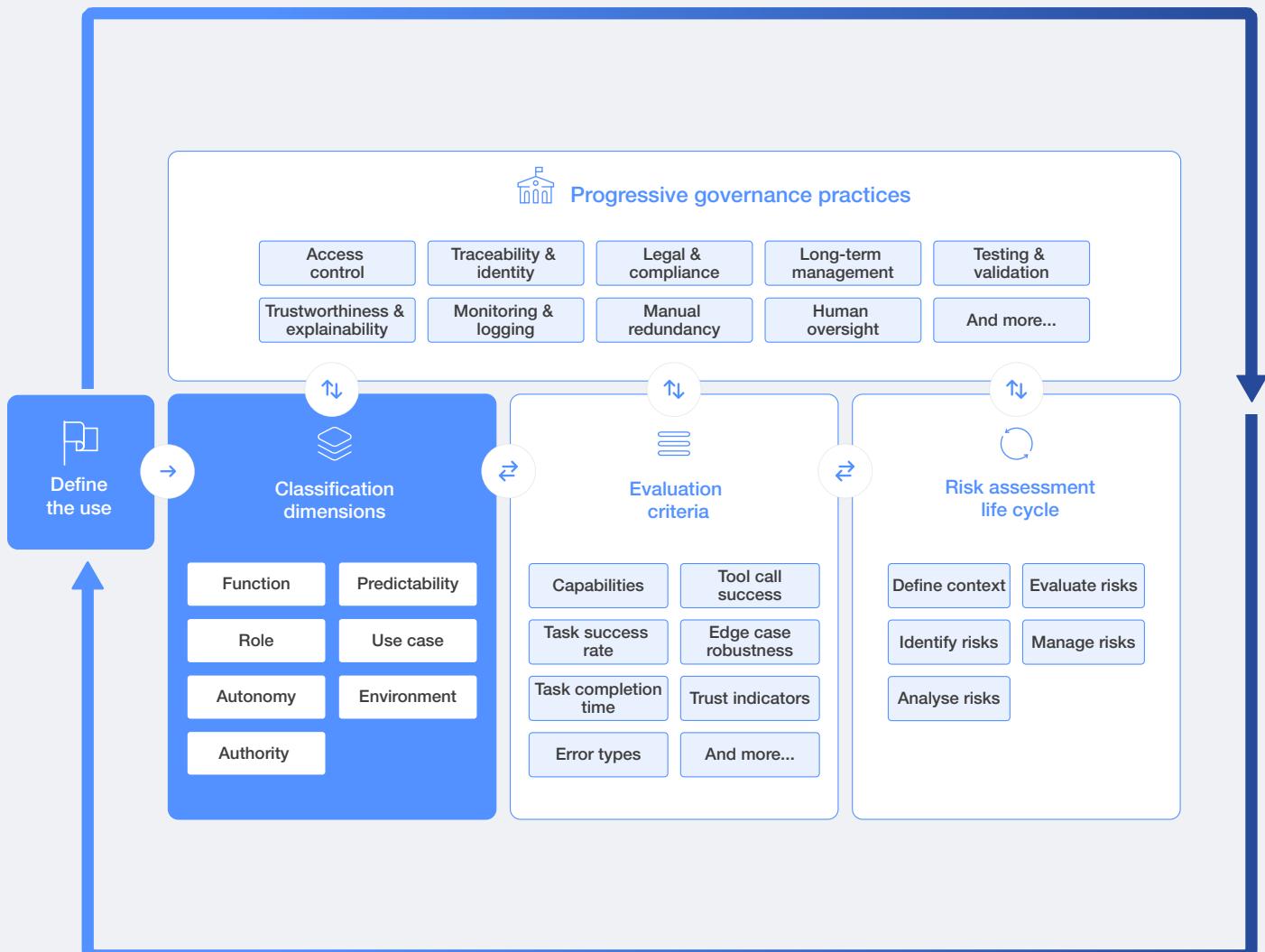
– **Guide governance and oversight:** Align safeguards, controls and monitoring mechanisms with the nature and complexity of the agent's role.

– **Support interoperability and scaling:** Structure agent types in ways that facilitate coordination in multi-agent environments and integration across systems.

Without clear classification, organizations may adopt AI agents without fully understanding what they are designed to do, how they operate, the impact they may have on their environment or the oversight mechanisms they require. This lack of clarity could result in gaps in safety, security, control, privacy, reliability and accountability.

FIGURE 7

Foundations for AI agent evaluation and governance – classification dimensions





2.2 Evaluation

Robust evaluation is crucial for assessing agent performance and limitations across diverse contexts.

As organizations begin deploying agents with different functional roles, the need for structured evaluation becomes more important. This section explores how evaluation methodologies are evolving to reflect this growing complexity.

Agent “evaluation” refers to the measurement of an AI agent’s performance and operation in representative contexts, generating evidence about how well it achieves intended functions, under what conditions and with what limitations. This means that robust evaluation frameworks are essential for building trust in AI agents’ performance. By providing clear, multidimensional assessments of agent capabilities and limitations, evaluations can help organizations develop appropriate expectations and confidence in agentic systems.

While the evaluation of foundation models such as LLMs is supported by a rich landscape of standardized benchmarks,^{18,19,20} agent evaluation remains nascent. Unlike static model testing, agents operate as orchestrated systems that combine tool use, memory, decision-making and user interaction, which exceed the scope of traditional benchmarks. In response, several agent-specific capability benchmarks have begun to emerge:

- **AgentBench:** Tests agents in interactive environments like web browsing and games, and is useful for evaluating real-time decision-making and adaptability²¹
- **SWE-bench:** Evaluates an agent’s ability to resolve GitHub issues in open-source repositories,

providing real-world measures of reasoning, code modification and system integration²²

- **HCAST:** Compares agent performance to human developers in areas such as programming tasks, offering calibrated insights into agent coding capabilities, for example²³

Although these emerging benchmarks offer valuable signals, they are typically built for academic or research settings, where tasks are predefined, environments are static and outcomes are often deterministic. They rarely capture operational realities such as ambiguous success criteria or dynamic workflows.

Evaluation requires clear performance metrics that capture both task-level and system-level outcomes. Examples include task success rate, completion time, error types, tool call success, throughput, robustness against edge cases and user trust indicators. These metrics help establish whether the system delivers its functions reliably and provide the operational evidence that later informs risk assessment and governance decisions.

Providers benchmark systems to assess technical maturity, while procurers and deployers are responsible for ensuring that agents operate safely and compliantly within specific industry, organizational and operational contexts. Therefore, deployment environments provide the most accurate ground truth, but deployers often lack the resources to design comprehensive benchmarks. In many cases, this makes collaboration with providers essential to establishing meaningful metrics.

An effective provider-focused evaluation should begin with a technical screening of baseline capabilities,

An effective provider-focused evaluation should begin with a technical screening of baseline capabilities, such as reasoning, planning and tool use.

such as reasoning, planning and tool use. Once validated in sandbox environments that mirror real-world tasks, agents may progress to controlled deployment, where they are integrated into workflows under close monitoring, with safeguards in place to confirm that they align with human or established decisions. Full deployment should only follow once reliability has been demonstrated, with fallback mechanisms and defined human oversight. Audit logs are central throughout this life cycle, providing structured records of agent activity and the rationale behind it. Audit logs also support governance by enabling oversight and accountability, aiding debugging by tracing errors and points of failure, and helping inform evaluation.

The following principles support this life cycle of agent evaluation:

- **Contextualization:** Reflect the tools, workflows and edge cases the agent will encounter in practice.
- **Multidimensional assessment:** Define success across various factors, including accuracy, robustness, latency tolerance, compliance, and user trust.
- **Temporal and behavioural monitoring:** Track performance over time to detect regressions, shifts in behaviour, or failures to adapt to evolving inputs.

Emerging evaluation tools are increasingly applied in enterprise settings to support the continuous assessment of agentic systems, helping to track reasoning, compare outcomes to expectations and detect anomalies that are overlooked by traditional testing. Major cloud providers have also started embedding such frameworks into their AI platforms, highlighting the importance of deployer-side evaluation for adoption.

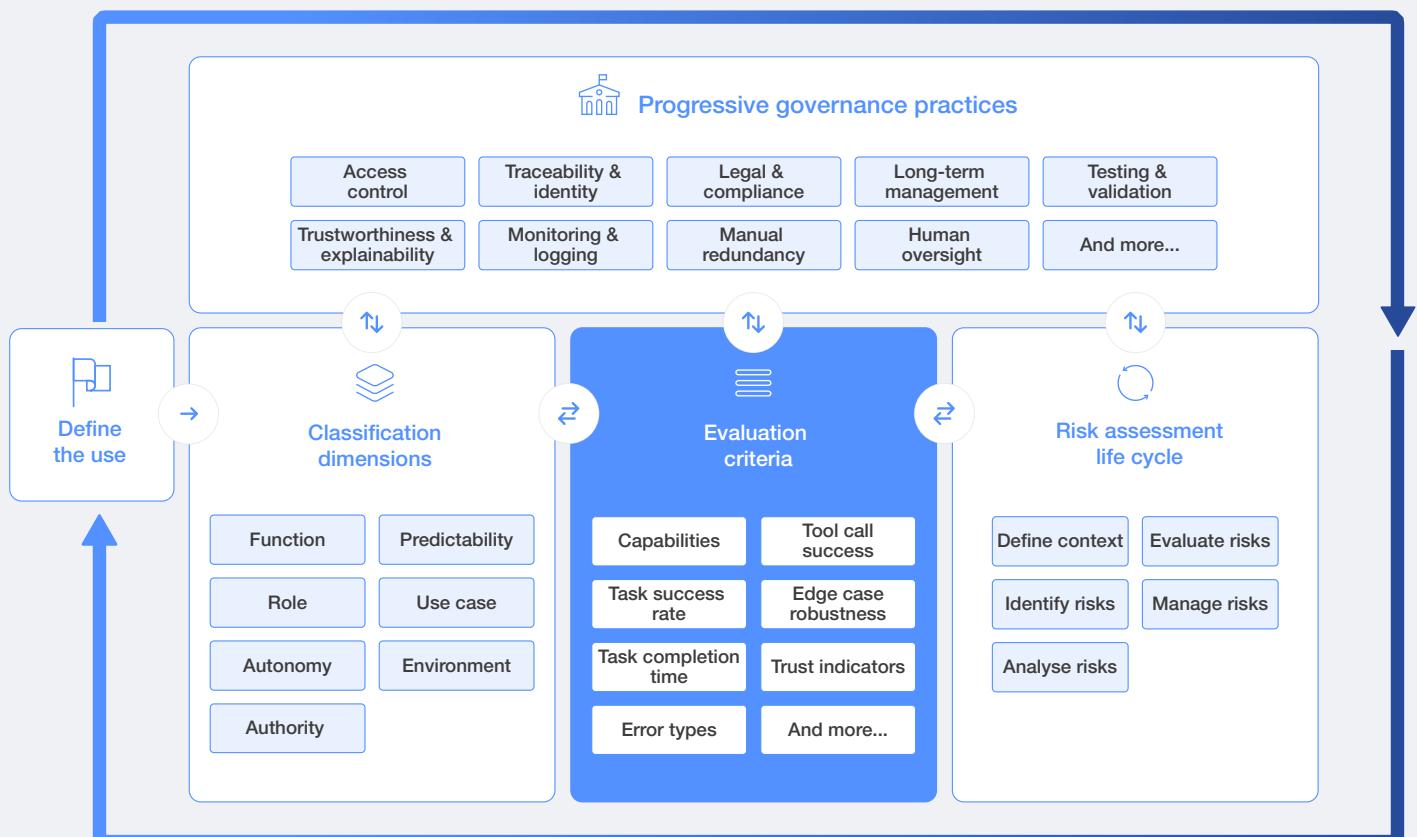
By approaching evaluation as a structured, context-aware and continuous process, organizations can more effectively determine whether an agent is fit for deployment.

To illustrate how these principles apply in practice, the following illustration examines a coding co-pilot agent. The illustration applies the evaluation dimensions from a deployer's perspective, showing how task-level and system-level metrics can be used to assess reliability, safety and overall performance in an operational setting.

Effective evaluation depends on close collaboration between providers and adopters, where transparent documentation, model specifications and performance reports from providers enable deployers to validate reliability, identify risks and apply safeguards throughout the system life cycle.

The results form an integrated performance profile that informs subsequent risk assessment and governance.

FIGURE 8 Foundations for AI agent evaluation and governance – evaluation criteria



```

    _object = mirror_obj
    if operation == "MIRROR_X":
        mod.use_x = True
        mod.use_y = False
        mod.use_z = False
    if operation == "MIRROR_Y":
        mod.use_x = False
        mod.use_y = True
        mod.use_z = False
    if operation == "MIRROR_Z":
        mod.use_x = False
        mod.use_y = False
        mod.use_z = True

    # selection at the end - add back the deselected
    if ob.select == 1:
        ob.select=1
    context.scene.objects.active = modifier_obj
    context.scene.objects.active.name = "selected" + str(modifier_obj)
    modifier_obj.select = 0
    bpy.context.selected_objects[0] = bpy.context.selected_objects[1]
    bpy.context.objects[one.name].select = 1

```

CASE STUDY 2

Coding co-pilot – evaluation

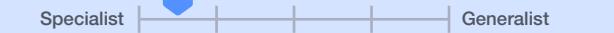
Coding co-pilot

Agent characteristics

1. Function

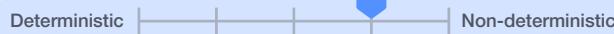
Assists human developers with code generation and debugging

2. Role



Specialist | Generalist

3. Predictability



Deterministic | Non-deterministic

4. Autonomy



Low | High

5. Authority



Low | High

Operational context

6. Use case

A coding co-pilot operates in the software development domain, assisting programmers within their coding environment by generating, completing and debugging code to improve productivity and reduce errors.

7. Environment



Simple | Complex

Coding co-pilot – evaluation

Evaluation starts with controlled tests in development environments to verify productivity gains while ensuring safety, reliability and compliance. Evaluation follows several key steps including:

- **Contextualization:** Testing across coding tasks such as code generation, debugging and documentation to reflect real workflows
- **Performance:** Measuring task success rate, completion time and error frequency, along with system metrics like tool-call success

- **Robustness:** Exposing the agent to ambiguous or conflicting code to assess recovery, error handling and adaptability
- **Human trust:** Gathering user feedback on reliability and usefulness
- **Monitoring:** Using continuous logging to detect performance drift, anomalous tool use or regressions after deployment

2.3 Risk assessment

Risk assessment identifies and analyses potential harms, linking evaluation results to oversight.

Evaluation establishes how the system performs, whereas risk assessment determines whether the agent and its use present risks that need to be understood, assessed and mitigated. Evaluation provides evidence as to whether the set mitigations are effective and met in implementation.

The goal of risk²⁴ assessment is to identify, analyse and prioritize the ways an agent could fail or be misused, estimate likelihood and severity, and determine whether it can operate within acceptable boundaries with appropriate controls. This applies to single agents and multi-agent systems, software-based and embodied deployments, and covers both technical and organizational vulnerabilities.

Risk assessment draws on an agent's defined classification dimensions to identify and analyse potential risks, considering factors such as cybersecurity threats, safety hazards, operational vulnerabilities, legal and regulatory requirements, and stakeholder impacts. It also incorporates evidence from evaluation activities, such as sandbox testing and pilot deployments, including task success rates, error patterns and robustness.

To make this process operational, organizations can follow a five-step life cycle that can be scaled to the complexity of the use case.

The life cycle outlined in Figure 9 links the outputs of classification and evaluation directly to risk management and progressive governance practices. The following, Table 1, provides an example of how the risk assessment process can be structured in practice.²⁵

FIGURE 9

Foundations for AI agent evaluation and governance – risk assessment life cycle

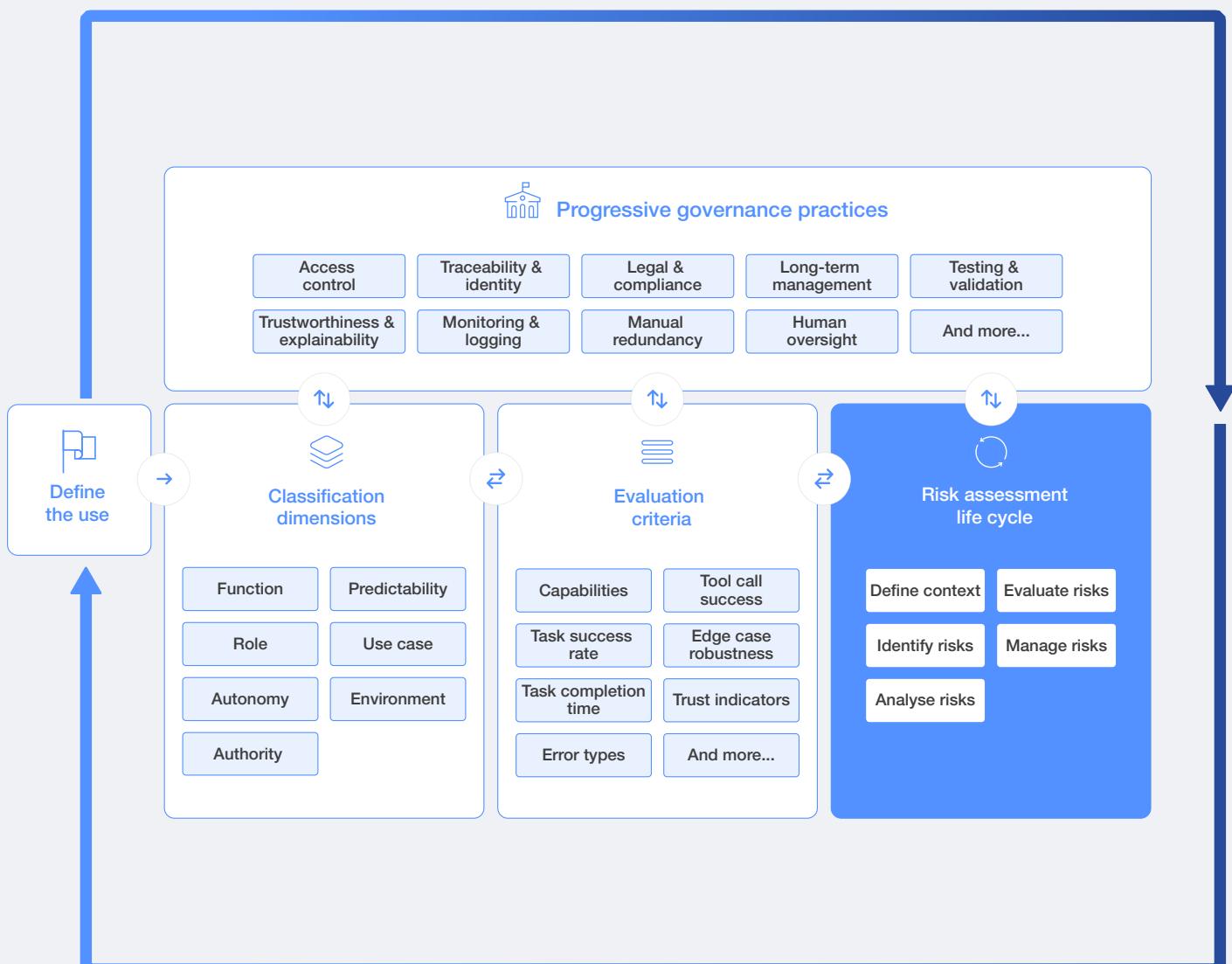


TABLE 1 Risk assessment life cycle for AI agents

 Step	 Objective	 Example activities	 Example outputs
1. Define context	Establish the scope of the assessment, system boundaries, objectives and criteria for managing risk	<ul style="list-style-type: none"> – Determine internal and external context (strategic goals, legal framework, stakeholders) – Define boundaries, intended use, assumptions – Establish risk criteria (likelihood, impact scales, acceptance threshold) 	<ul style="list-style-type: none"> – Context definition – Risk management plan – Risk evaluation criteria
2. Identify risks	Identify potential technical, organizational and ecosystem risks, harms and affected parties	Brainstorm, workshops, risk identification (e.g. hazard identification, threat identification, etc.), identification of sources of risk, causes, failure mode analysis	<ul style="list-style-type: none"> – Risk register listing risks, causes, impacts
3. Analyse risks	Understand the nature, likelihood and consequence of each risk and quantify them	<ul style="list-style-type: none"> – Assess probability and impact (considering, for example, characteristics like autonomy and authority, predictability and operational context) – Identify existing controls or guardrails; apply qualitative or quantitative methods for risk estimation; use evaluation results to inform likelihood and impact 	<ul style="list-style-type: none"> – Risk analysis scores showing likelihood impact ratings and rationale
4. Evaluate risks	Compare analysis results with risk criteria to determine priority and tolerability	<ul style="list-style-type: none"> – Rank and prioritize risks – Use evaluation results for quantifying and prioritizing risks – Use performance metrics and test confidence to inform risk thresholds 	<ul style="list-style-type: none"> – Risk ranking summary – Risk acceptance evaluations
5. Manage risks	Implement risk response actions (avoid, mitigate, transfer, accept) and monitor risks	<ul style="list-style-type: none"> – Assign owners of preventive, detective and response controls – Evidence these controls through evaluation results – Address emerging risks as systems evolve or context changes – Integrate feedback loops for continuing monitoring – Coordinate incident response and impact mitigation – Update governance and controls based on lessons learned 	<ul style="list-style-type: none"> – Control actions – Implementation plan – Residual risk profile – Risk assessment report – Evidence logs – Monitoring reports – Revised frameworks – Improved processes
<p>Defining clear risk criteria and tolerability thresholds, and applying them consistently to prioritize and evaluate risks, remains a central challenge in AI risk management.</p> <p>The identification, analysis and evaluation of risks are directly linked to the classification dimensions introduced earlier, allowing organizations to understand how factors such as autonomy, authority, predictability and environmental complexity shape overall risk levels for AI agents. Inherent risk combines likelihood and impact, while residual risk reflects the effectiveness of applied mitigations, informed by evaluation evidence such as system</p>		<p>reliability, robustness and observed error rates. This relationship establishes a clear connection between how an agent is designed, how it performs and how risks are managed, providing the basis for proportionate governance and oversight.</p> <p>Applying this approach in practice helps demonstrate how structured risk assessment translates classification and evaluation evidence into measurable controls. The following example illustrates the risk assessment process in the context of an autonomous vehicle.</p>	



CASE STUDY 3

Autonomous vehicle – risk assessment

Autonomous vehicle

Agent characteristics

1. Function

Performs the complete driving task without human control

2. Role

Specialist → Generalist

3. Predictability

Deterministic → Non-deterministic

4. Autonomy

Low → High

5. Authority

Low → High

Operational context

6. Use case

An autonomous vehicle operates in the transportation domain, navigating public or private road environments to transport passengers or goods safely and efficiently without direct human control.

7. Environment

Simple → Complex

Autonomous vehicle – risk assessment

Risk assessment focuses on identifying and mitigating possible failures across perception, decision-making and control systems. Key risk areas include sensor malfunction, data drift, adversarial interference and coordination failures with other vehicles or infrastructure that could lead, for example, to loss of steering or braking control and eventual collisions.

Each risk is analysed for its likelihood (for example, the frequency of sensor failure leading to braking failure) and its impact (for example, the severity of injury, fatality or legal consequence).

Quantitative scoring combines these factors and is weighted according to the vehicle's autonomy and authority levels.

Mitigation measures may include redundancy and diversity in critical sensors, reduction of autonomy or authority thresholds, anomaly detection mechanisms and real-time incident reporting. Residual risk is evaluated after these safeguards are applied, drawing on evidence from controlled testing, field trials and continuous monitoring. The results determine whether the system can safely progress to wider deployment or requires additional control layers.

 **Governance levels are informed by risk assessment outcomes, ensuring that controls scale with demonstrated autonomy, authority and contextual complexity.**

2.4 Governance considerations for AI agents: a progressive approach

Progressive governance approaches scale oversight and safeguards in proportion to the autonomy, authority and complexity of the agent.

Evaluation and risk assessment provide critical insights into an agent's capabilities, performance, reliability, security, safety and alignment. Governance, however, determines whether those insights translate into effective oversight and responsible adoption. "Governance" refers to the structured application of technical safeguards and operational, ethical and organizational processes intended to ensure agents remain within acceptable risk boundaries over time. As agents become more capable and integrated into core workflows, governance must evolve from basic precautionary measures to dynamic, multi-layered systems of control and accountability. Governance levels are informed by risk assessment outcomes, ensuring that controls scale with demonstrated autonomy, authority and contextual complexity.

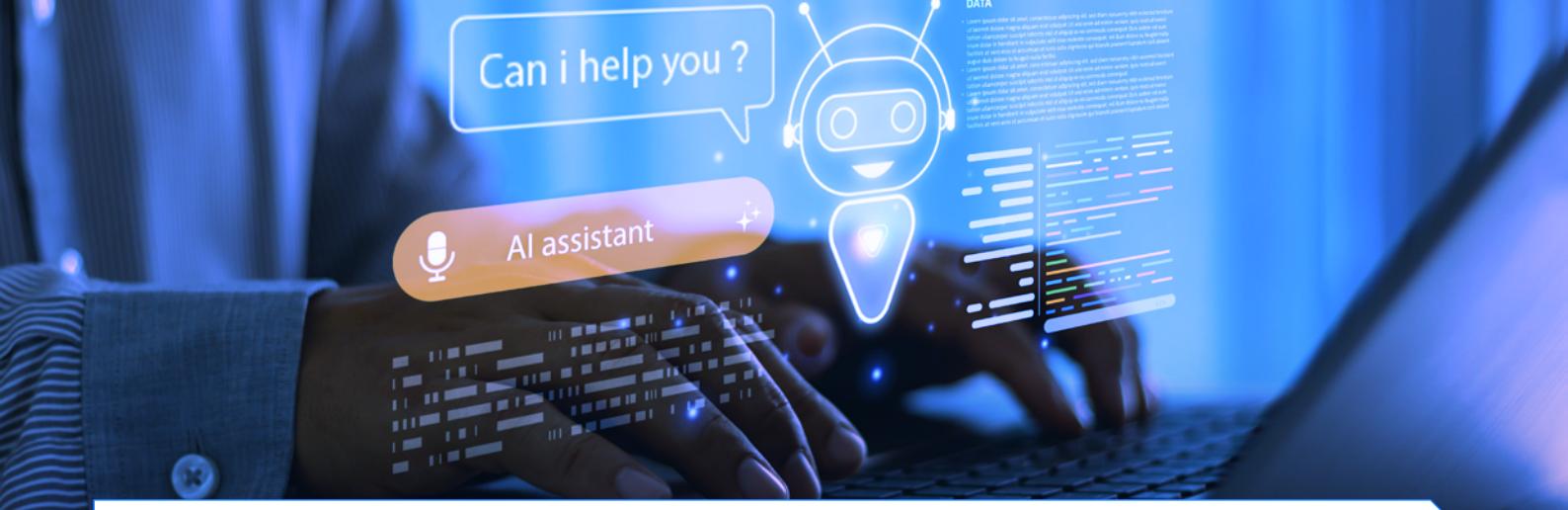
A progressive set of governance levels can be distinguished, ranging from baseline safeguards to enhanced controls and systemic risk management. These levels correspond to the agent's classification profile, which is linked to its function, predictability, autonomy, authority and operational context. Oversight, therefore, intensifies as agents move from narrow, low-risk applications to complex, high-impact environments.

a control plan with clear ownership and verification and validation steps, operating limits and monitoring requirements, and a deployment status. These outputs feed directly into progressive governance, ensuring oversight scales in line with an agent's demonstrated risk profile and operating context.

Across these levels, governance mechanisms advance in both scope and sophistication. The focus shifts from operational safeguards to comprehensive risk management, with early levels emphasizing reactive measures, while more advanced levels incorporate proactive monitoring, accountability frameworks and systemic risk assessments.

This progression is evident across key areas such as monitoring, accountability, risk management, transparency, adaptability and scope. Monitoring evolves from basic logging to real-time, AI-assisted oversight, incorporating the automated analysis of logs to detect anomalies and deviations in system behaviour. In parallel, risk management advances from static checklists to dynamic, predictive modelling, while the scope of governance expands from narrow, task-specific oversight to consideration of broader ecosystem impacts.

Operational environments are dynamic, and effective governance often requires recalibrating autonomy and authority in real time. The following example illustrates this through a personal assistant agent, whose level of autonomy and authority is dynamically adjusted to ensure ongoing compliance.



CASE STUDY 4

Personal assistant – governance considerations

Agent characteristics

Agent characteristics

1. Function

Assists users by organizing schedules, managing communication and coordinating

2. Role

Specialist |————| Generalist

3. Predictability

Deterministic |————| Non-deterministic

4. Autonomy

Low |————| High

5. Authority

Low |————| High

Operational context

6. Use case

It operates in the personal productivity domain, managing tasks, communications and information across a user's digital environment to support daily coordination and decision-making.

7. Environment

Simple |————| Complex

Personal assistant – governance considerations

Governance focuses on scaling oversight in line with the personal assistant's autonomy, authority and environmental complexity. Unlike narrow task agents, a personal assistant operates across multiple platforms such as email, calendars, messaging and enterprise tools, raising questions about the extent of information it can access, interpret and act upon on behalf of the user. As integration deepens and authority expands (e.g. from drafting messages to sending them, or booking travel) governance mechanisms must increase.

Key governance risks include data overreach, privacy violations, prompt manipulation and unauthorized actions such as unintended communication.

Mitigation measures include least-privilege access, consent-based data sharing, input and output filtering, audit logging, and human approval for sensitive actions. Adaptive controls should reduce permissions upon detecting anomalies or policy breaches, supported by continuous monitoring and incident reporting.

The example illustrates that an agent's overall impact emerges from the interaction of multiple dimensions across function, role, predictability, autonomy, authority and context. As these dimensions shift, so does the risk profile, reinforcing the need for governance frameworks that are both progressive and adaptive.

Effective governance requires maintaining an appropriate level of human oversight in relation to the agent's autonomy, authority and operational context. In high-risk or less predictable settings,

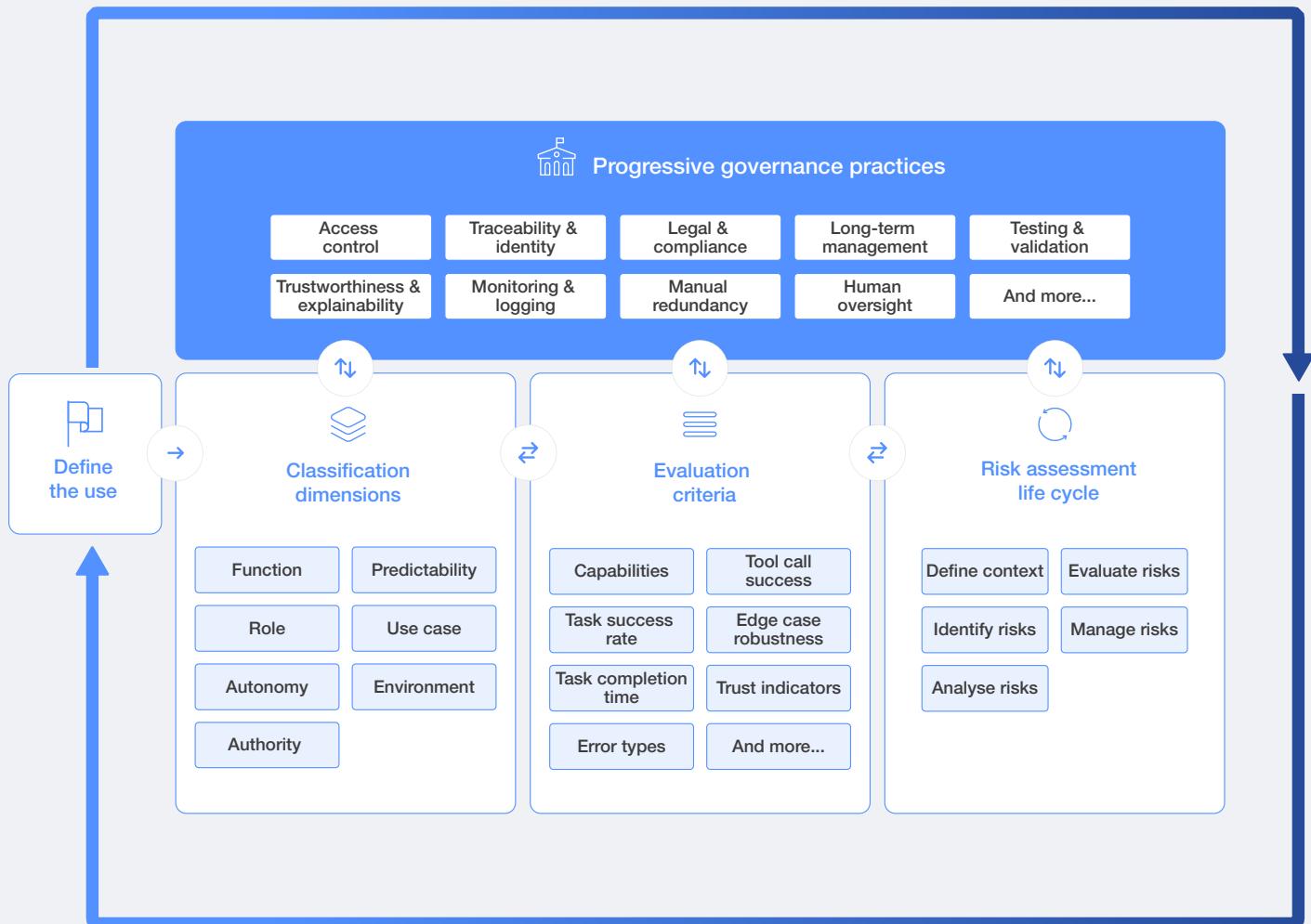
a human-in-the-loop (HITL) configuration ensures that agents can suggest or prepare actions, but final decisions remain subject to explicit human approval. In more stable or clearly defined environments, a human-on-the-loop (HOTL) configuration allows agents to act within defined boundaries, while humans monitor behaviour, receive alerts and retain the ability to intervene or override when necessary. Integrating these oversight models into governance structures helps maintain accountability and human judgment as agents operate with greater independence and scale.

TABLE 2

Baseline governance mechanisms for AI agents

Governance area	Foundational mechanism	Purpose
 Access control	Enforce least-privilege access; define task boundaries.	Prevent each agent from accessing unnecessary data, systems, or tools; reduce risk of misuse or accidental harm.
 Legal and compliance	Conduct a data protection impact assessment (DPIA); perform privacy and regulation compliance checks, such as General Data Protection Regulation or the California Consumer Privacy Act (CCPA).	Ensure data handling and processing complies with relevant laws and regulations.
 Testing and validation	Perform sandbox runs or controlled pilots with non-production data; install input-output filters; perform third-party audits.	Validate expected behaviour, detect errors and prevent untested code from affecting live systems, conduct audits (code, red teaming, etc.).
 Monitoring and logging	Implement logging for all agent actions; set up anomaly alerts or dashboards.	Maintain traceability for accountability; enable early detection, incident response and post-incident analysis.
 Human oversight	Define and assign oversight models, including HITL/HOTL. Require policy review before deployment and set supervisory triggers for exceptions.	Ensure accountable human control for material decisions, keep behaviour aligned with organizational policies and provide escalation paths when the agent acts unexpectedly.
 Traceability and identity	Assign unique agent identifiers; tag outputs to the responsible agent instance.	Attribute actions and outcomes to specific agents; enable forensic review and performance tracking.
 Long-term management	Establish protocols for ongoing monitoring, updates and eventual decommissioning.	Ensure continued alignment, performance and relevance throughout the agent's life cycle.
 Trustworthiness and explainability	Implement explainability tools; establish trust metrics.	Ensure agent behaviour is interpretable and measurable; build user confidence.
 Manual redundancy	Establish manual redundancy procedures to ensure the sustained continuity of critical business use cases.	Preserve data integrity and plan for human resources to take over.

FIGURE 10 | Foundations for AI agent evaluation and governance: progressive governance practices



“Prior to deployment, agents should undergo sandbox or controlled pilot testing using non-production data to validate expected behaviour.

For all agents, regardless of their level of autonomy, authority or the complexity of their operational context, specific governance mechanisms should serve as a baseline for adoption. At a minimum, every agent should operate under strict access control based on the principle of least privilege, with clear task boundaries that prevent unnecessary system or data access. Basic legal and compliance checks, such as data protection impact assessments and privacy compliance reviews, are necessary to ensure alignment with regulatory obligations. In addition, technical controls such as input and output filters can help constrain agent behaviour by screening potentially harmful, irrelevant or non-compliant interactions before they propagate through the system.

Prior to deployment, agents should undergo sandbox or controlled pilot testing using non-production data to validate expected behaviour and mitigate unintended effects. All actions and planning should be recorded in an audit log for traceability, supported by monitoring tools or alerts tailored to the agent's overall profile. This enables

the detection of anomalies early, while balancing concerns about privacy and surveillance risks associated with monitoring at scale. Human oversight, through policy reviews, audit log analysis and supervisory triggers, helps ensure alignment with organizational priorities. Unique identifiers and output tagging support attribution, performance tracking and post-incident analysis. In practice, the depth of safeguards should scale with the agent's autonomy, authority, complexity of context and overall impact. Higher-risk systems require proportionally greater investment in monitoring and oversight, with a deliberate balance between human review and automated, continuous monitoring.

By embedding these measures into the life cycle of all agents, organizations establish a governance baseline that can scale proportionally with complexity and risk. This foundation helps address immediate operational safety and compliance needs, creating the structures and practices upon which more advanced, context-specific governance mechanisms can be layered as agents become more autonomous, integrated and capable.

Looking ahead: multi-agent ecosystems

Future ecosystems of interacting agents introduce new risks that demand interoperable standards and oversight.

 As organizations begin to deploy multiple agents across departments, systems and networks, a new class of failure modes is emerging.

Future ecosystems of interacting agents introduce new risks that demand interoperable standards and oversight.

The future of AI agents will happen in a much broader space than enterprise automation and will increasingly be defined by the emergence of multi-agent ecosystems. In these ecosystems, agents are expected to interact, negotiate and collaborate across organizational and technical boundaries. In many ways, the interconnectedness of these systems will redefine the future of AI, moving beyond traditional enterprise automation to allow agents to negotiate, collaborate and coordinate autonomously. While this shift opens new opportunities for innovation, it also introduces challenges around alignment, trust, emergent behaviours and system design. Given the complex nature of these systems, ensuring responsible behaviour and effective use requires robust mechanisms for monitoring and assessing agent interactions. A few examples of emerging multi-agent ecosystems and their implications are:

- **Agent-to-agent commerce:** Agents can initiate transactions, request services or exchange data with other agents, forming a new layer of internet activity with considerable downstream economic implications.
- **Internet of agents:** Beyond isolated interactions, large-scale networks of agents could form an “internet of agents,” raising questions of interoperability, standards, governance and societal impact.
- **Trust frameworks for inter-agent collaboration:** As agents begin operating autonomously across boundaries, establishing shared norms, credentialing systems and behavioural standards is critical to verify identity, capabilities and reliability.
- **Agent governance and oversight:** As agent capabilities advance, dedicated “governor” or “auditor” agents will monitor, audit or regulate the actions of other agents, validating transactions, detecting anomalies and correcting unsafe or unintended behaviours. They enable scalable oversight in complex ecosystems, but they risk overreliance on agents supervising other agents.

- **Embodied agents:** Embodied agents extend governance challenges into the physical world, where oversight mechanisms must address both digital actions and consider physical safety, reliability and human interaction.

As organizations begin to deploy multiple agents across departments, systems and networks, a new class of failure modes is emerging, linked to potentially misaligned interactions between agents. A few examples include:

- **Orchestration drift:** When agents are plugged into other agents without shared context or coordination logic, workflows can become brittle or unpredictable.
- **Semantic misalignment:** When two agents interpret the same instruction differently, it can lead to conflicting actions or duplicated effort, with implications for safety, reliability and coordination.
- **Security and trust gaps:** Without shared trust frameworks, agents may inadvertently expose sensitive data or interact with malicious actors, exploiting vulnerabilities in the system.
- **Interconnectedness and cascading effects:** Failures in tightly linked agents or systems can propagate across networks, creating a chain of disruptions.
- **Systemic complexity:** As the number and diversity of interacting agents grow, the likelihood of emergent behaviours and cascading failures increases, making them more difficult to anticipate, trace or diagnose.

Although the widespread deployment of multi-agent ecosystems is still in its early stages, providers and adopters must now anticipate the associated risks. As organizations experiment and pilot agents, misaligned interactions are already creating new failure modes. Understanding possible challenges such as orchestration drift, semantic misalignment and cascading failures enables adopters to implement safeguards before scaling. A proactive approach ensures responsible growth, aligning governance with technical capabilities and defined boundaries.

Conclusion

Agents have already begun moving into production across various domains, including customer support, workflow automation, autonomous research and more. As adoption advances and as early use cases move from single agents to more complex interconnected systems, expectations for scalable oversight grow.

This report has outlined the foundations for AI agent evaluation and governance, presenting a conceptual approach to classification, evaluation, risk assessment and governance that supports responsible adoption. The proposed dimensions aim to help organizations better understand what an agent does, how it operates and its place within the broader organization. Evaluation provides evidence of performance and reliability, while risk assessment identifies potential harms and mitigations. Governance helps translate these insights into safeguards and concrete accountability mechanisms, which can then scale as the agent's capability is extended to more complex use-cases and scenarios.

As the development of agents advances towards multi-agent ecosystems, the need for shared protocols, interoperability standards and coordinated oversight is only going to increase. Cross-functional governance that links technical assurance with organizational accountability is considered key to preventing cascading failures and ensuring responsible oversight at scale.

At the core of this long-term transition is effective human-AI collaboration. In evolving governance practices, clear responsibility for objectives, supervision, and outcomes must be supported by novel tools and processes that maintain systems as understandable, safe and secure in practice.

Ultimately, the responsible deployment of agentic systems depends on a baseline of trust, transparency and accountability that remains valid for all digital systems. With thoughtful design, careful evaluation and proportionate governance, AI agents are likely to amplify human capabilities, improve productivity and, over time, meaningfully contribute to both public and private value.

Contributors

The World Economic Forum's AI Governance Alliance Safe Systems and Technologies working group convenes chief science officers and AI producers to advance thought leadership surrounding AI agents, from their architecture to applications, social implications, guardrails and governance structures. This initiative promotes the development of safety mechanisms and encourages collaboration on best practices for the design and implementation of AI systems.

World Economic Forum

Benjamin Cedric Larsen

Initiatives Lead, AI Safety, Centre for AI Excellence

Capgemini

Olivier Denti

Data Architect, AI, Capgemini Invent

Jason DePerro

Human-AI Collaboration Director, Capgemini Invent

Jeanne Heure

Vice-President, Digital Trust & Security, Capgemini Invent

Raymond Millward

GenAI for R&D Technical Solution Lead, Capgemini Engineering

Efi Raili

Safety Authority, Technology and Innovation, Capgemini Engineering

Acknowledgements

Animashree (Anima) Anandkumar

Bren Professor of Computing and Mathematical Sciences, California Institute of Technology (Caltech)

Mandanna Appanderanda Nanaiah

Head, Infosys Responsible AI, North America, Infosys

Nebahat Arslan

Director, Group General Counsel and Partnership Officer, Women in AI

Mennatallah El-Assady

Professor of Interactive Visualization and Intelligence Augmentation, ETH Zurich

Ricardo Baeza-Yates

WASP Professor, KTH Royal Institute of Technology, Sweden

Amir Banifatemi

Chief Responsible AI Officer, Cognizant

William Bartholomew

Director of Public Policy, Responsible AI, Microsoft

Aaron Bawcom

Field Chief Technology Officer, Invisible Technologies

Pete Bernard

Chief Executive Officer, EDGE AI FOUNDATION

Fabio Casati

Lead, AI Trust and Governance Lab, ServiceNow

Kevin Chung

Chief Strategy Officer, Writer

Cathy Cobey

Global Trusted AI Advisory Leader, EY

Ben Colman

Co-Founder and Chief Executive Officer, Reality Defender

Sakyasingha Dasgupta

Founder and Chief Executive Officer, EdgeCortix

Umeshwar Dayal

Senior Fellow and Senior Vice-President, Hitachi America; Corporate Chief Scientist, Hitachi

Mona Diab

Director, Language Technologies Institute, Carnegie Mellon University

Yawen Duan

AI Safety Research Manager, Concordia AI

Gilles Fayad

Adviser, Institute of Electrical and Electronics Engineers (IEEE)

Claudia Fischer

Public Policy Planning, Global Affairs, OpenAI

Jenn Gamble

Head, Data Science, Distly AI

Chen Goldberg Senior Vice-President, Engineering, CoreWeave	Mao Matsumoto Head, NEC Fellow Office, NEC
Tom Gruber Founder, Humanistic AI	Sean McGregor Agentic Product Safety Lead, MLCommons
Gillian Hadfield Professor of Law and Professor of Strategic Management, University of Toronto	Risto Miikkulainen Professor of Computer Science, The University of Texas at Austin
Peter Hallinan Director, Responsible Artificial Intelligence, Amazon Web Services (AWS)	Satwik Mishra Executive Director, Centre for Trustworthy Technology (CTT)
Bennet Hillenbrand Agentic Product Safety Lead, MLCommons	Margaret Mitchell Researcher and Chief Ethics Scientist, Hugging Face
Babak Hodjat Chief AI Officer, Cognizant	Jessica Newman Director, AI Security Initiative, Centre for Long-Term Cybersecurity, UC Berkeley
Sean Kask Chief AI Strategy Officer, SAP	Mark Nitzberg Executive Director, Center for Human-Compatible AI, UC Berkeley
Robert Katz Vice-President, Responsible AI and Tech, Salesforce	Henrik Ohlsson Vice-President; Chief Data Scientist, C3 AI
Michael Kearns Founding Director, Warren Center for Network and Data Sciences, University of Pennsylvania	Dmytro Ovcharenko AI Chief Technology Officer, Ministry of Digital Transformation of Ukraine
Steven Kelly Chief Trust Officer, Institute for Security and Technology	Maria Pocovi Global Head of Responsible AI, Uniphore
Alex Lebrun Co-Founder and Chief Executive Officer, Nabla	Reza Rooholamini Chief Scientific, Artificial Intelligence and Innovation Officer, CCC Intelligent Solutions
Stefan Leichenauer Vice-President, Engineering, SandboxAQ	Long Ruan Chief Technology Officer, Astra Tech
Tze Yun Leong Professor of Computer Science, National University of Singapore	Jason Ruger Chief Information Security Officer, Lenovo
Scott Likens Global AI and Innovation Technology Lead, PwC	Daniela Rus Director, Computer Science and Artificial Intelligence Laboratory (CSAIL), Massachusetts Institute of Technology (MIT)
Ramana Lokanathan Senior Vice-President, Engineering and AI, Automation Anywhere	Jun Seita Team Director, Medical Science Deep Learning Team, RIKEN
Nada Madkour Non-Resident Research Fellow, University of California, Berkeley	Norihiro Suzuki Chairman of the Board, Hitachi Research Institute, Hitachi
Richard Mallah Principal AI Safety Strategist, Future of Life Institute	Sumit Taneja Senior Vice-President and Global Head, Artificial Intelligence (AI) Consulting and Implementation, EXL Service
Pilar Manchón Senior Director, Engineering, Google	
Gaonyalelwe Maribe Head, Data Analytics and AI, Old Mutual	
Darko Matovski Founder and Chief Executive Officer, causaLens	

Fabian Theis	Science Director, Helmholtz Association
Li Tieyan	Chief AI Security Scientist, Huawei Technologies
Lisa Titus	AI Policy Manager, Meta
Kush Varshney	IBM Fellow, IBM
Anthony Vetro	President, Chief Executive Officer, IEEE Fellow, Mitsubishi Electric Research Laboratories
Tiffany Wang Xingyu	Founder, Stealth
Andrea Wong	Global Head, Responsible AI Policy, Trust and Safety, Bytedance
Lauren Woodman	Chief Executive Officer, DataKind
Michael Young	Vice-President of Products, Private AI
Xiaohui Yuan	Director, Innovation Research Center; Senior Expert, TRI, Tencent Holdings
Andy Zhang	Researcher, Stanford University
Leonid Zhukov	Vice-President of Data Science, Boston Consulting Group X (BCG X); Director, BCG Global AI Institute, Boston Consulting Group (BCG)

World Economic Forum

Abhi Balakrishnan	Initiatives Lead, AI and Innovation, Centre for AI Excellence
Maria Basso	Head, AI Applications and Impact, Centre for AI Excellence
Daniel Dobrygowski	Head, Governance and Trust, Centre for AI Excellence
Audrey Duet	Head, Data and AI Innovation, Centre for AI Excellence
Ginelle Greene	Initiatives Lead, Artificial Intelligence and Energy, Centre for AI Excellence
Connie Kuang	Initiatives Lead, Technology Convergence, Centre for AI Excellence
Cathy Li	Head, Centre for AI Excellence; Member of the Executive Committee
Hesham Zafar	Lead, Partner Engagement, Centre for AI Excellence
Production	
Laurence Denmark	Creative Director, Studio Miko
Blake Elsey	Designer, Studio Miko
Will Liley	Editor, Studio Miko

Endnotes

1. Capgemini Research Institute. (2024). *Harnessing the value of generative AI*. <https://www.capgemini.com/wp-content/uploads/2024/05/Final-Web-Version-Report-Gen-AI-in-Organization-Refresh.pdf>.
2. Organisation for Economic Co-operation and Development (OECD). (2024). *Recommendation of the Council on Artificial Intelligence*. <https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449>.
3. National Institute of Standards and Technology (NIST). (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=936225.
4. International Organization for Standardization (ISO). (2023). *ISO/IEC 23894:2023: Information technology — Artificial intelligence — Guidance on risk management*. <https://www.iso.org/standard/77304.html>.
5. Claude Docs. (n.d.). Features overview. <https://docs.claude.com/en/docs/build-with-claude/overview>.
6. Anthropic. (2024). *Introducing the Model Context Protocol*. <https://www.anthropic.com/news/model-context-protocol>.
7. Surapaneni, R., M. Jha, M. Vakoc and T. Segal. (2025). *Announcing the Agent2Agent Protocol (A2A)*. Google for Developers. <https://developers.googleblog.com/en/a2a-a-new-era-of-agent-interoperability/>.
8. Mitchell, M., S. Wu, A. Zaldivar, P. Barnes, et al. (2019). Model Cards for Model Reporting. *FAT* 19: Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 220-229. <https://dl.acm.org/doi/10.1145/3287560.3287596>.
9. Parikh, S. and R. Surapaneni. (2025). *Powering AI commerce with the new Agent Payments Protocol (AP2)*. Google Cloud. <https://cloud.google.com/blog/products/ai-machine-learning/announcing-agents-to-payments-ap2-protocol>.
10. Cloudflare. (n.d.). *Zero Trust security | What is a Zero Trust network?* <https://www.cloudflare.com/en-gb/learning/security/glossary/what-is-zero-trust/>.
11. Hasan, M. M., L. Hao, E. Fallahzadeh, B. Adams, et al. (2025). *Model Context Protocol (MCP) at First Glance: Studying the Security and Maintainability of MCP Servers*. <https://arxiv.org/abs/2506.13538v1>.
12. Lynch, B. and R. Harang. (2025). *From Prompts to Pwns: Exploiting and Securing AI Agents*. <https://i.blackhat.com/BH-USA-25/Presentations/US-25-Lynch-From-Prompts-to-Pwns.pdf>.
13. Adapted from: International Organization for Standardization (ISO). (2023). *ISO/IEC 42001:2023: Information Technology — Artificial intelligence — Management system*. <https://www.iso.org/standard/81230.html>; National Institute of Standards and Technology (NIST). (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=936225.
14. Ibid.
15. Capgemini. (n.d.). *Business, meet agentic AI*. https://www.capgemini.com/wp-content/uploads/2025/05/Confidence-in-autonomous-and-agentic-systems_19May.pdf.
16. SAE International. (2021). *J3016_202104 - Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*. https://www.sae.org/standards/j3016_202104-taxonomy-definitions-terms-related-driving-automation-systems-road-motor-vehicles.
17. Russell, S. J. and P. Norvig. (2021). *Artificial Intelligence: A Modern Approach*. Pearson.
18. Hendrycks, D., C. Burns, S. Basart, A. Zou, et al. (2021). *Measuring Massive Multitask Language Understanding*. <https://arxiv.org/abs/2009.03300>.
19. Srivastava, A., A. Rastogi, A. Rao, A. A. Shoeb, et al. (2022). *Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models*. Transactions on Machine Learning Research (TMLR). <https://arxiv.org/abs/2206.04615>.
20. Liang, P., R. Bommasani, T. Lee, D. Tsipras, et al. (2022). *Holistic Evaluation of Language Models*. Transactions on Machine Learning Research (TMLR). <https://arxiv.org/abs/2211.09110>.
21. Liu, X., H. Yu, H. Zhang, Y. Xu, et al. (2024). *AgentBench: Evaluating LLMs as Agents*. International Conference on Learning Representations (ICLR). <https://arxiv.org/abs/2308.03688>.
22. Jimenez, C. E., J. Yang, A. Wettig, S. Yao, et al. (2023). *SWE-bench: Can Language Models Resolve Real-World GitHub Issues?* <https://arxiv.org/abs/2310.06770>.
23. Rein, D., J. Becker, A. Deng, S. Nix, et al. (2025). *HCAST: Human-Calibrated Autonomy Software Tasks*. <https://arxiv.org/abs/2503.17354>.
24. "Risk" refers to the composite measure of an event's probability (or likelihood) of occurring and the magnitude or degree of the consequences of the corresponding event; National Institute of Standards and Technology (NIST). (2024). *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*. <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf>.
25. Adapted from: International Organization for Standardization (ISO). (2023). *ISO/IEC 42001:2023: Information Technology — Artificial intelligence — Management system*. <https://www.iso.org/standard/81230.html>; National Institute of Standards and Technology (NIST). (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=936225.



COMMITTED TO
IMPROVING THE STATE
OF THE WORLD

The World Economic Forum, committed to improving the state of the world, is the International Organization for Public-Private Cooperation.

The Forum engages the foremost political, business and other leaders of society to shape global, regional and industry agendas.

World Economic Forum
91–93 route de la Capite
CH-1223 Cologny/Geneva
Switzerland

Tel.: +41 (0) 22 869 1212
Fax: +41 (0) 22 786 2744
contact@weforum.org
www.weforum.org